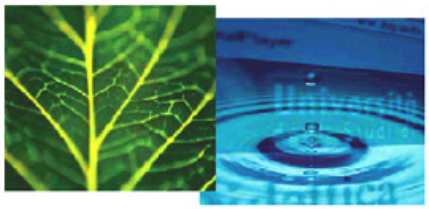
PhD Dissertation



International Doctorate School in Information and Communication Technologies

DISI - University of Trento

MINING HUMAN BEHAVIORS: AUTOMATED BEHAVIORAL ANALYSIS FROM SMALL TO BIG DATA

Jacopo Staiano

Advisor:

Prof. Nicu Sebe

Università degli Studi di Trento

28th of April, 2014

Nothing is mysterious, no human relation. Except love.
Susan Sontag[1]

Quando penso a Pasolini,
a come agiva rispetto alla società,
alle cose, mi stimo molto poco.
Massimo Troisi

[1] As Consciousness Is Harnessed to Flesh: Journals and Notebooks, 1964-1980.

Acknowledgements

First, I want to thank Nicu for all the support and for believing in me (sometimes even more than I myself did). It was, is, and will be, a great privilege to work under your guidance, to witness the birth and growth of the M-HUG group, and to have you as a great friend. I thank Prof. Theo Gevers, and the friends and former colleagues at the Intelligent Systems Lab of the University of Amsterdam: my experience working there is what actually brought me to push a PhD. Vladimir Nedovic, Roberto Valenti, Ivo Everts, Jose Alvarez and Hamdi Dibeklioglu deserve a special mention.

I am grateful to Prof. Hamid Aghajan at Stanford, for having me as a visiting researcher in the Ambient Intelligence Lab at Stanford. To the 1737 crew, Scott Zimmerman, Handan Selcuk, Sujay and Naveem Vennam, Trip, Cemre and Yesim Ozkurt, Giuseppe "tac" and Flo Valente: no matter how far, we're connected.

Thanks to Prof. Alex "Sandy" Pentland, for giving me the opportunity of working at the Human Dynamics Lab at MIT in 2011 and 2013, and to Nuria Oliver for having me as intern at Telefonica I+D in Barcellona.

I finally want to thank: Bruno Lepri and Fabio Pianesi for the fruitful (and ongoing) collaborations, Marco Guerini, Stefano Teso, Andrea Passerini, Leo Maccari, Elena Pavan, Elisa Ricci, Gabriele Catania, Alex Cappelletti, Maria Menendez, Antonella De Angeli, Matteo Bonifacio, Andrey Bogomolov along with colleagues at DISI and all my coauthors; Gloria, Jasper, Vika, Ram, Julian, and the entire M-HUG group; Manuel, Francesca, Andrea, Mario and Danilo at DISI; the friends at Mozart (RIP), Cafe' de la Paix, Chinaski, Qubé Café, RockInCapri, PistoiaBlues, NoGuRu; Anna & Mario at Castelli Romani; Riccardo Esposito at My Social Web; all lifelong friends DOP, Antonio "azazel", Antonio "hope", Antonio "goodbirds", Vincenzo "the shell", Cecco, Graziano, Marcella, Gabriella, Luigi, Ottavio, Colonnesi, Fausto, Marialaura, Cekkini, Palillo, Catuogno, the Guardino Bros.

Last but not least, I am deeply grateful to my parents and family for their endless support and patience. And to Ilaria, for what we are and for what we'll be.

List of Publications

- J Staiano, N Oliver, B Lepri, R de Oliveira, M Caraviello, N Sebe
  Money Walks: a Human-Centric Study on the Economics of Personal Mobile Information [258]
  Proceedings of the 2014 ACM Conference on Ubiquitous Computing, ACM UBICOMP 2014;

- J Staiano, M Guerini
  Depeche Mood: a Lexicon for Emotion Analysis from Crowd Annotated News [254]
  The 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014;

- R Subramanian, Y Yan, J Staiano, O Lanz, N Sebe
  On the relationship between head pose, social attention and personality prediction for unstructured and dynamic group interactions [268]
  15th International conference on multimodal interaction, ACM ICMI 2013;

- S Teso, J Staiano, B Lepri, A Passerini, F Pianesi
  Ego-centric graphlets for personality and affective states recognition [276]
  International Conference on Social Computing, 874-877, IEEE SocialCom 2013;

- M Guerini, J Staiano, D Albanese
  Exploring Image Virality in Google Plus [105]
  International Conference on Social Computing, 874-877, IEEE SocialCom 2013;

- MK Abadi, J Staiano, A Cappelletti, M Zancanaro, N Sebe
  Multimodal Engagement Classification for Affective Cinema [2]
  Humaine Association Conference on Affective Computing and Intelligent Interaction, IEEE ACII 2013;

- G Zen, N Rostamzadeh, J Staiano, E Ricci, N Sebe
  Enhanced semantic descriptors for functional scene categorization [317]
  21st International Conference on Pattern Recognition, ICPR 2012;

- J Staiano, B Lepri, N Aharony, F Pianesi, N Sebe, A Pentland
  Friends don't lie: inferring personality traits from social network structure [255]
  Proceedings of the 2012 ACM Conference on Ubiquitous Computing, 321-330, ACM UBICOMP 2012;

- B Lepri, J Staiano, G Rigato, K Kalimeri, A Finnerty, F Pianesi, N Sebe, A Pentland
  The sociometric badges corpus: A multilevel behavioral dataset for social behavior in complex organizations [165]
  International Conference on Social Computing, 623-628, IEEE SocialCom 2012;

- B Lepri, R Subramanian, K Kalimeri, J Staiano, F Pianesi, N Sebe
  Connecting Meeting Behavior with Extraversion – A Systematic Study [167]
  IEEE Transactions on Affective Computing, 2012;

- J Staiano, M Menéndez, A Battocchi, A De Angeli, N Sebe
  UX_Mate: From facial expressions to UX evaluation [257]
  Proceedings of the Designing Interactive Systems Conference, 741-750, ACM DIS 2012;

- J Staiano, B Lepri, R Subramanian, N Sebe, F Pianesi
  Automatic modeling of personality states in small group interactions [256]
  Proceedings of the 19th ACM international conference on Multimedia, 989-992, ACM MM 2011;

- H Joho, J Staiano, N Sebe, JM Jose
  Looking at the viewer: analysing facial activity to detect personal highlights of multimedia contents [137]
  Multimedia Tools and Applications, 51(2):505-523, 2011;

- B Lepri, R Subramanian, K Kalimeri, J Staiano, F Pianesi, N Sebe
  Employing social gaze and speaking activity for automatic determination of the extraversion trait [166]
  International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction, ACM ICMI 2010;

- R Subramanian, J Staiano, K Kalimeri, N Sebe, F Pianesi
  Putting the pieces together: multimodal analysis of social attention in meetings [267]
  Proceedings of the international conference on Multimedia, 659-662, ACM MM 2010.

Abstract

This research thesis aims to address complex problems in Human Behavior Understanding from a computational standpoint: to develop novel methods for enabling machines to capture not only what their sensors are perceiving but also how and why the situation they are presented with is evolving in a certain manner.

Touching several fields, from Computer Vision to Social Psychology through Natural Language Processing and Data Mining, we will move from more to less constrained scenarios, describing models for automated behavioral analysis in different contexts: from the individual perspective, e.g. a user interacting with technology, to the group perspective, e.g. a brainstorming session; from living labs, e.g. hundreds of people transparently tracked in their everyday life through smart-phone sensors, to the World Wide Web.

Contents

Chapter 1

Introduction

Humans are social by nature, machines are asocial by design. To bridge this gap, it is desirable to build systems able to correctly interpret social interactions between humans, and thus exploit forms of *Automatic Human Behavior Understanding*. As computing becomes ubiquitous and the number of available sensors increases, researchers can mine and exploit the huge amount of behavioral data produced by people in their everyday lives, in order to build behavioral models and gain valuable insights from collective and individual perspectives.

This thesis elaborates methods for *Automatic Human Behavior Understanding* in two interaction scenarios: the first, "classic", scenario including situations in which sensor(s) and subject(s) share the same physical space (*e.g.* a user interacting with a machine, as well as subjects participating in a meeting, having a coffee break, etc); the second, extending the former by exploiting situations in which people interact by means of ubiquitous devices: the sensors lie in the physical dimension of each individual engaging into forms of physical and/or virtual interaction with other people.

The behavior of human beings is influenced by both subjective and objective factors. The former, in this thesis referred to as *internal* determinants of behavior, include characteristics such as emotional state and personality: *"what I do depends on how I feel"*; conversely, the latter refer to factors external to the

subject, that is characteristic of the specific situation the subject finds herself in: *"what I do depends on what is happening around me"*.

Previous research has extensively shown that such internal determinants of behavior, or feelings, are often manifested through very short and unconscious body movements: from facial expressions to back-channeling gestures (e.g. fidgeting), we all use such *social signals* to interpret others' behavior and tune, in turn, ours. We thus apply computer vision and speech processing techniques to automatically detect such events on which our behavioral models are built.

Furthermore, drawing from extensive literature in Social Psychology, where *personality* is recognized as being a causal determinant of people's behavior, our research focuses on building personality based behavioral models. Such models will ultimately allow design and deployment of proactive systems in heterogeneous contexts, including, but not limited to, Surveillance, Health-Care, User eXperience, Human Resources Selection, Tutoring, Targeted Marketing.

1.1 The Context

Aristotle's definition of humans as *social animals* has persisted during the centuries and entered as common knowledge: such *companionable animals*, as Dante Alighieri calls them, are characterized by an innate tendency (what Karl Marx referred to as *Gemeinwesen* [186]) to form groups, interact with the environment, collaborate toward reaching common goals and improve their community's conditions. Exploiting recent advances in the fields of Computer Vision, Machine Learning, Natural Language and Speech Processing, Network and Data Science, along with the increase in computational power of modern microprocessors, a novel and promiscuous research field has been lately emerging and gaining interest among researchers: *Computational Social Science*, leveraging *the capacity to collect and analyze data with unprecedented breadth, depth and scale* [163].

This research thesis aims to address complex problems in *Human Behavior Understanding* from a computational standpoint: to develop novel methods for enabling machines to capture not only *what* their sensors are perceiving but also *how* and *why* the situation they are presented with is evolving in a certain manner. We endorse a interactionist [171, 90] approach to human behavior analysis, considering behavior as a function of both the person and his/her environment; in other terms, we embrace both the situationist theory of "people responding to an environment that consists of other people responding to *their* environment, which consists of people responding to an environment of people's responses" [243], and the classical person-perspective which sees individual characteristics as principal determinants of behavior.

In this framework, internal determinants of behavior such as personality *traits*, psycho-pathological risk factors, and other individual predispositions, are treated as dynamically evolving attributes under the influence of external determinants such as situational and structural attributes of the environment a person is behaving *within*.

Personality is an individual's characteristic comprising all attributes (behavioral, temperamental, emotional, and mental) affecting his/her dispositions. In our everyday lives, we describe people as being more or less *sociable*, *talkative*, or bold: we constantly employ these descriptors to explain and/or predict others' behavior, attaching them to well-known and new acquaintances. *Extroversion*, the personality *trait* they refer to, is so familiar that we exploit it continuously in a natural and inconspicuous manner. Similarly, we talk about other people being more/less prone to frustration and anger (*Neuroticism/Emotional Stability*), responsible or attentive (*Conscientiousness*), and so on. Being *personality* recognized as an individual characteristic representing the subjective tendency to react emotionally or behaviorally in a certain way, if these traits do manifest themselves through the perceivable physical behaviors of the subject, then they can indeed be automatically assessed. In fact, recent works have validated such

hypothesis [14, 166, 182, 183, 204, 205, 222].

Recently, a dynamic view of personality traits has been proposed [84], suggesting that traits can be reconstructed through density distributions of personality *states*. In other words, personality states can be seen as specific behavioral episodes in which a subject behaves more or less extrovertly, neurotically, and so on.

Thus, in order to capture, interpret, and/or induce change to, behaviors of people or groups of, it seems necessary to identify important constituents of behavior such as personality *states* and, at the same time, recognize how change in the environment and such constituents mutually influence each other.
Depending on scale and context of the environment in which a system of interactions is taking place, different strategies can be adopted; expert humans are, in fact, trained to exploit such internal determinants by driving people in certain directions and analysing their micro-behaviors: influential recent works [80, 219] provide evidence of hard-to-fake perceptual cues that can be systematically treated, hence potentially exploitable by a machine.

Consider, for instance, a standard human resources selection task: the recruiters, after screening the applicants' curricula and selecting a dozen of them, set a meeting in their offices. Once the candidates are gathered around a table, the recruiters drive them into predefined discussions to analyze their behaviors while engaging with others, and to derive information on their psycho-attitudinal profiles. Finally, the candidates are singularly interviewed. Since the biggest source of information for the recruiter is the non-verbal behavior of the subject, we can postulate that a machine equipped with the necessary sensors (such as webcams and microphones), and trained for the task, might be able to exploit the very same cues in order to select the best applicant for the position.

Furthermore, the wide deployment of mobile devices and increase in broadband connectivity have in the last decade re-shaped the intrinsic working of society. The network of interactions between people has become more dense,

by orders of magnitude, maximizing reach and information flow. Such reshaping brought dramatic positive effects (e.g. innovation, faster response to events, etc.) along with side-effects such as information overload, and pervasive surveillance. Everyone of us, in this very moment, is producing valuable data through a smart-phone (by all means, a *personal tracking device*). Providers of software, connectivity, and some times even hardware are creating value from personal data, following a business model based on data centralization and separation on the one hand, and customer profiling and targeted advertising on the other. The result, nowadays, is that big silos of personal information have been built and monetized, in a centralized fashion, by those who control the infrastructure used to gather, store, and mine such data.

In Chapters 4 and 5 we present a few works that show how informative such *digital breadcrumbs* can be, while in Chapter 6 we will elaborate on the world-changing impact technologies built for cooperation in a decentralized fashion can have.

1.2 Structure of the Thesis

This thesis is structured as follows: in Chapter 2 we report the state of the art and present our contributions in modeling human behavior within the Human Computer Interaction (HCI) scenario; in Chapter 3 we describe relevant and related works dealing with human behavior understanding in social contexts, and detail our work and the advancements obtained; in Chapter 4 we enlarge the scope of our analyses to the least controlled scenarios, using wearable sensors and smartphones as primary sensing infrastructure to gather information about collective and individual behaviors.

Finally, in Chapter 5 we discuss works that analyse and exploit behavioral manifestations in the World Wide Web, and elaborate future research directions in Chapter 6.

More specifically, this thesis develops and is organized under the metaphor of a magnifying glass: we start at the individual scale, focusing on the analysis of a single person interacting with technology [257] (Section 2.2) or enjoying multimedia content [137] (Section 2.1); then, we zooms out, widening the field of view to include more people physically interacting with each other in scenarios with varying constraints: specifically a meeting room first [267, 256] (Sections 3.1 and 3.2) , a whole building [165, 276] (Sections 4.2 and 4.3), and communities of people transparently sensed through their own smartphone devices [255] (Sections 4.1 and 4.4); in the latter stage, further widening our scope, all connections to the physical world are abandoned from the sensing point of view and behavioral dynamics appearing in a purely virtual domain such as internet are analyzed [105] (Sections 5.1 and 5.2).

Thus, at each zoom out step, behavioral modeling is applied to increasingly large groups of people in settings with decreasing levels of control.

This thesis consists of the following publications:

- Chapter 2:
  - H Joho, J Staiano, N Sebe, JM Jose
    Looking at the viewer: analysing facial activity to detect personal highlights of multimedia contents [137]
    Multimedia Tools and Applications, 51(2):505-523, 2011;
  - J Staiano, M Menéndez, A Battocchi, A De Angeli, N Sebe
    UX_Mate: From facial expressions to UX evaluation [257]
    Proceedings of the Designing Interactive Systems Conference, 741-750, ACM DIS 2012.
- Chapter 3:
  - R Subramanian, J Staiano, K Kalimeri, N Sebe, F Pianesi
    Putting the pieces together: multimodal analysis of social attention in

meetings [267]
Proceedings of the international conference on Multimedia, 659-662, ACM MM 2010;

– J Staiano, B Lepri, R Subramanian, N Sebe, F Pianesi
Automatic modeling of personality states in small group interactions [256]
Proceedings of the 19th ACM international conference on Multimedia, 989-992, ACM MM 2011.

- Chapter 4:

  – J Staiano, B Lepri, N Aharony, F Pianesi, N Sebe, A Pentland
  Friends don't lie: inferring personality traits from social network structure [255]
  Proceedings of the 2012 ACM Conference on Ubiquitous Computing, 321-330, ACM UBICOMP 2012;

  – B Lepri, J Staiano, G Rigato, K Kalimeri, A Finnerty, F Pianesi, N Sebe, A Pentland
  The sociometric badges corpus: A multilevel behavioral dataset for social behavior in complex organizations [165]
  International Conference on Social Computing, 623-628, IEEE SocialCom 2012;

  – S Teso, J Staiano, B Lepri, A Passerini, F Pianesi
  Ego-centric graphlets for personality and affective states recognition [276]
  International Conference on Social Computing, 874-877, IEEE SocialCom 2013;

  – J Staiano, N Oliver, B Lepri, R de Oliveira, M Caraviello, N Sebe
  Money Walks: a Human-Centric Study on the Economics of Personal Mobile Information [258]
  Proceedings of the 2014 ACM Conference on Ubiquitous Computing,

ACM UBICOMP 2014[1].

- Chapter 5:
  - M Guerini, J Staiano, D Albanese
    Exploring Image Virality in Google Plus [105]
    International Conference on Social Computing, 874-877, IEEE SocialCom 2013;
  - J Staiano, M Guerini
    Depeche Mood: a Lexicon for Emotion Analysis from Crowd Annotated News [254]
    (The 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014).

The following papers have been published during the course of the Ph.D. but are not included in this thesis:

- B Lepri, R Subramanian, K Kalimeri, J Staiano, F Pianesi, N Sebe
  Employing social gaze and speaking activity for automatic determination of the extraversion trait [166]
  International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction, ACM ICMI 2010;

- B Lepri, R Subramanian, K Kalimeri, J Staiano, F Pianesi, N Sebe
  Connecting Meeting Behavior with Extraversion – A Systematic Study [167]
  IEEE Transactions on Affective Computing, 2012;

- G Zen, N Rostamzadeh, J Staiano, E Ricci, N Sebe
  Enhanced semantic descriptors for functional scene categorization [317]
  21st International Conference on Pattern Recognition, ICPR 2012;

[1] work performed at Telefonica I+D, Barcellona

- MK Abadi, J Staiano, A Cappelletti, M Zancanaro, N Sebe
  Multimodal Engagement Classification for Affective Cinema [2]
  Humaine Association Conference on Affective Computing and Intelligent Interaction, IEEE ACII 2013;

- R Subramanian, Y Yan, J Staiano, O Lanz, N Sebe
  On the relationship between head pose, social attention and personality prediction for unstructured and dynamic group interactions [268]
  15th International conference on multimodal interaction, ACM ICMI 2013.

Chapter 2

Deriving and Exploiting Behavioral Insights at Individual Level

In this chapter, we report two original deployments of web-cam based sensing systems and their validation:

- an approach to detecting personal highlights in videos based on an analysis of facial activities;
- an ecologically valid vision-based system for automatic User eXperience (UX) evaluation.

First, we present an approach to detecting personal highlights in videos based on an analysis of facial activities. Our facial activity analysis was based on the motion vectors tracked on twelve points in human face. The magnitude of the motion vectors represented a degree of a viewer's affective reaction to video contents in our approach. We examined 80 facial activity videos recorded for 10 participants, each watched eight video clips in various genres. The experimental results suggest that useful motion vectors to detect personal highlights varied significantly across viewers. However, it was suggested that the activity in the upper part of face tended to be more indicative of personal highlights than the activity in the lower part.

Secondly, we propose and evaluate UX_Mate, a non-invasive system for the automatic assessment of user experience (UX). In addition, we contribute a novel database of annotated and synchronized videos of interactive behavior and facial expressions.

UX_Mate is a modular system which tracks facial expressions of users, interprets them based on pre-set rules, and generates predictions about the occurrence of a target emotional state, which can be linked to interaction events. The system simplifies UX evaluation providing an indication of event occurrence. UX_Mate has several advantages compared to other state of the art systems: easy deployment in the user's natural environment, avoidance of invasive devices, and extreme cost reduction. We report a pilot and a validation study on a total of 46 users, where UX_Mate was used for identifying interaction difficulties.

The studies show encouraging results that open possibilities for automatic real-time UX evaluation in ecological environments.

2.1 Analysing Facial Activities to Detect Personal Highlights of Videos

The explosion of multimedia contents and the need for effective access have resulted in the development of a number of video summarisation techniques. Video summaries are needed in many situations. For example, such a summary could be useful for getting a gist of the video content. Summaries can also support the end-user's decision-making to view the entire video (e.g., films) or not. The results of such decision making can then be used for modelling the user preference. This also suggests that the techniques developed for video summarisation can be related to a task of user profiling and/or personal recommendation of unseen videos.

Money and Agius [195] categorise video summaries based on three dimen-

sions: content type (feature based, object based, event based, and perception based), personalisation (personalised, generic), and interactivity (interactive, static). Techniques such as shot boundary detection and keyframe extraction are the basis of the feature based summaries which have been extensively investigated [111]. This type of summaries is not designed to consider semantics of video contents. The summaries investigated in evaluation forums such as TRECVID [210] tend to be object based or event based summaries. Such a summary consists of unique scenes of an object such as "antique car" or an object in the context of an event "red hot air balloon ascending". These types of summaries are designed to present a gist of contents based on the main objects and events within a video. However, the feature based and object/event based approaches tend to suffer from the semantic gap problem in interactive use of such summaries.

Recently, there has been a growing interest in perception based summaries. These look at a higher level of abstraction than the other types of summaries by exploiting viewer's affective state, perceived excitement, and attention found within or caused by video contents [196], [136]. Perception based approaches are designed to overcome the semantic gap problem in summarisation by finding affective scenes in videos. Another prospect of the perception based summarisation is the application of creating the personalised summaries. Since the affective scenes in videos are subjective, and hence, can vary across viewers, personalised summaries that are tailored to one's preference can be generated from the same video. However, this area has not been fully exploited, and existing techniques to generate perception based summaries are expensive. For example, they require manual annotations [277] or several physiological sensors [196, 197] to capture people's affective state.

In this work, we distinguish the terms *facial activity* and *facial expression*. The facial activity refers to the movement of specific points (i.e., motion vectors) in human face, while the facial expression refers to a category of human

emotion inferred from a classification of multiple motion vectors. The two concepts will be explained in more detail later.

The rest of this section is structured as follow. We first review the work on affective video analysis and summarisation. Then we briefly present the facial expression recognition system. The data collection method and evaluation measurement are then described, followed by the results of analysis and discussion. We conclude the paper by discussing some directions of future work.

2.1.1 Affective Video Summarisation

Annotation according to affective or emotional categories of video is a relatively young domain, gaining more and more importance [37, 110, 112, 145, 194, 295, 314]. The main objective is to make the recommendation personalized and situation sensitive. If the affective content of a video is detected, it will be very easy to build an intelligent video recommendation system, which can recommend videos to users based on users' current emotion and interest. For example, when the user is sad, the system will automatically recommend happy movies to him/her; when the user is tired, the system may suggest a relaxing movie.

All the current affective analysis systems try to solve the following problems [295]: 1) identification of valid affective features; 2) bridging the gap between affective features and affective states; 3) establishing an affective model to take user's personality into consideration; 4) representing the affective state.

In general, there are three kinds of popular affective analysis methods. Categorical Affective Content Analysis methods usually define a few basic affective groups and discrete emotions, for example, "happy", "sadness" and "fear". The videos or parts of them are classified automatically into one of these categories. Moncrieff et al. [194] analyze changes in sound energy of the non-literal components of the audio tracks of films and detect four sound energy events commonly used in horror films: "surprise or alarm", "apprehension or emphasis

of a significant event", "surprise followed by a sustained alarm", and "building apprehension up to a climax". They find that these four sound energy events convey well established meanings through their dynamics to portray and deliver certain affect or sentiment related to the horror film genre. Kang et al. [145] detect emotional events such as fear, sadness and joy from videos by computing intra-scene context (shots' coherences, shot's interactions, dominant features in color and motion information) and inter-scene context (scene's relationship with other scenes). Xu et al. [314] identify video/audio segments which make audience laugh in comedy and scary segments in horror films as affective contents. They use Hidden Markov Models (HMM) based audio classification method to detect audio emotional events (AEE) such as laughing, horror sounds, etc. Then, they use the AEE as a clue to locate the corresponding video segment.

The second type of affective analysis method is called Dimensional Affective Content Analysis method, which commonly employs the Dimensional Affective Model to compute the affective state. The psychological Arousal-Valence (A-V) Affective Model is one popular Dimensional Affective Model [71]. Arousal stands for the intensity of affective experience and Valence characterizes the level of "pleasure". Hanjalic and Xu [112] did research on affective state representation and modeling by using the A-V Affective Model. According to the A-V affective model, the affective video content can be represented as a set of points in the two-dimensional (2-D) emotion space that is characterized by the dimensions of arousal (intensity of affect) and valence (type of affect). By using the models that link the arousal and valence dimensions to low-level features extracted from video, the affective video content can be mapped onto the 2-D emotion space. Then, an affect curve (arousal and valence time curves) can be easily detected as a reliable representation of the expected transitions from one feeling to another along a video. Pleasure-Arousal- Dominance (P-A-D) model [191] is another popular affective model. Pleasure stands for the degree of pleasantness of the emotional experience, Arousal stands for the level

of activation of the emotion, and Dominance describes the level of attention or rejection of the emotion. Based on P-A-D model, Arifin et al. [16] propose to use Dynamic Bayesian Networks (DBNs) to build up a P-A-D value estimator, which estimates the P-A-D values of the video shots of the input video. Then, the video can be segmented based on the estimated P-A-D content. Different from the Arousal and Valence modeling proposed by Hanjalic and Xu [112], this work takes the influences of former emotional events and larger emotional events into consideration.

The third type of affective analysis method is Personalized Affective Content Analysis method. The representative work is reported in [295], which introduces more personalization factors into affective analysis and apply this to Music Video (MV) retrieval. First, they build a user interface and record the users' feedback in the user profile database. Each profile records MV's ID, user's descriptions about MV's Arousal and Valence (two scores describing their opinions about Arousal and Valence level). When users play MV, they can also use feedback to change their opinions on MV at any time. Based on the users' profile, two Support Vector Regression (SVR) models (Arousal model and Valence model) are trained to fit the user's affective descriptions. Finally, the affective features extracted from MV are fed into the trained models to get the personalized affective states. The authors also provide a novel Affective Visualization interface for efficient and user-friendly MV retrieval. Through this interface, the user can easily log into the system, search MV based on their affective states (for example, anger, happy, sad/blue, or peaceful) and also provide his/her feedback on each MV.

Directly relevant to our present work, Money and Agius [195] provide a taxonomy of video summaries and their generation techniques based on an extensive literature survey. We use their taxonomy to discuss existing work on video summarisation and relate our work to them.

The first aspect of their framework is the information sources analysed for

summarisation. *Internal* summarisation techniques analyse internal information from video streams produced during the production stage of video contents. More specifically, they tend to use low-level image, audio, and text features of videos. *External* summarisation techniques analyse external information which can be obtained from the process of capturing, producing, or viewing videos. External summarisation techniques are further divided into *User-based information* and *Contextual information* sources. User-based information typically includes people's behaviour during the interaction with video contents. This also includes people's preference information. The user-based information can be obtained in an *obtrusive* way using explicit feedback or in an *unobtrusive* way using various sensors. While unobtrusive methods are generally preferred, they tend to be noisy and limited in the level of details [195]. An example of the contextual information is the geographical footprints of videos using a GPS facility equipped with a video camera.

Both internal or external information have been exploited for affective video summarisation. The examples of internal information are Hanjalic and Xu [112] (discussed above) and Chan and Jones [41]. Chan and Jones [41] present a prototype system for affect-based indexing and retrieval of films, which is based on audio feature extraction. By analyzing all the audio data (speech, music, special effects and silence), the authors extracted the continuum of arousal and valence within the time dimension and used it to develop an affect annotation scheme.

The external information is often obtained by physiological sensors. For example, Mooney et al. [197] performed a preliminary study of the role of viewer's physiological states in an attempt to improve data indexing for search and within the search process itself. Participants' physiological responses to emotional stimuli were recorded using a range of biometric measurements, such as galvanic skin response (GSR), skin temperature, and other. The study provides some initial evidence that supports the use of biometrics as the user-based

external information. Soleymani, et al. [252] proposed a method for affective ranking of movie scenes, which takes into account both user emotions as well as video content. User emotion behaviour was inferred based on evidence gathered from the measurements of five peripheral physiological signals (galvanic skin response, electromyogram, blood pressure, respiration pattern and skin temperature), as well as self-assessments. In addition, the movie scenes were analysed using various video and audio features, which portrayed significant events within those scenes.

The approach investigated in this paper belongs to the group of Categorical Affective Analysis and can be seen as an *external* summarisation technique using *user-based* information. More specifically, we exploited viewer's facial expression while watching videos to find affective scenes for summarisation. Our information source (i.e., facial expression) was obtained in an *unobtrusive* way. This has a potential to make our approach simpler, more practical, and more feasible when compared to other approaches which exploited physiological signals of viewers. For example, in Money and Agius [196], subjects were wrapped by a sensor belt around their chest, a watch-type device was put around a wrist, and other signals were captured from several finger tips, and finally, their arm was rested on a cushion on the table. On the other hand, our approach required only a conventional web camera with which most recent PCs and laptops are equipped.

The next subsections describe our system and the method to generate affective summaries by exploiting viewer's facial expressions.

2.1.2 Facial Expression Recognition System

Our real time facial expression recognition system is composed of a face tracking algorithm which outputs a vector of motion features of certain regions of the face. The features are fed as inputs to a Bayesian network classifier. The system has been described in detail in [247] and for completeness we briefly

Figure 2.1: A snap shot of our realtime facial expression recognition system. On the left side is a wireframe model overlaid on a face being tracked. On the right side the correct expression, Angry, is detected.

describe the components of the system in the following sections. A snap shot of the system, with the face tracking and recognition result is shown in Figure 2.1.

Face and facial feature tracking

The face tracking technique used in our system is an improved version of the system developed by Tao and Huang [274] called the piecewise Bezier volume deformation (PBVD) tracker. Our face tracker uses a model-based approach where an explicit 3D wireframe model of the face is constructed (see Fig. 2.1). A generic face model is warped to fit the detected facial features. The face model consists of 16 surface patches embedded in Bezier volumes. The surface patches defined this way are guaranteed to be continuous and smooth.

Once the model is constructed and fitted, head motion and local deformations of the facial features such as the eyebrows, eyelids, and mouth can be tracked. First the 2D image motions are measured using template matching between frames at different resolutions. Image templates from the previous frame and from the very first frame are both used for more robust tracking. The measured

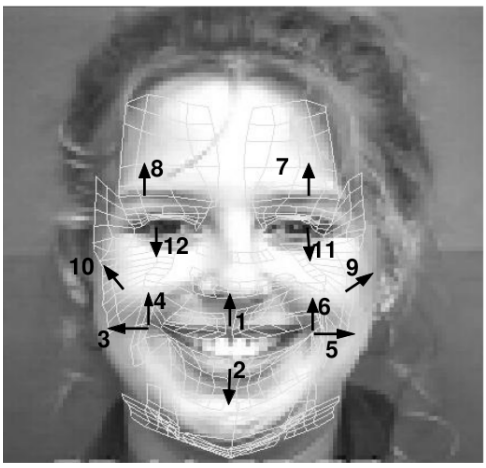


Figure 2.2: The facial motion measurements.

2D image motions are modelled as projections of the true 3D motions onto the image plane. From the 2D motions of many points on the mesh, the 3D motion can be estimated by solving an overdetermined system of equations of the projective motions in the least squared sense.

The recovered motions are represented in terms of magnitudes of some pre-defined motion of various facial features. Each feature motion corresponds to a simple deformation on the face, defined in terms of the Bezier volume control parameters. We refer to these motions vectors as Motion-Units (MUs). Note that they are similar but not equivalent to Ekman's AU's [79] and are numeric in nature, representing not only the activation of a facial region, but also the direction and intensity of the motion.

The 12 MUs used in the face tracker are shown in Fig. 2.2. As you can see, the first six vectors are roughly located in the lower part of human face while the other six vectors are located in the upper part of the face. We will denote the 12 MUs as $MU1$, $MU2$, ..., $MU12$ in this paper. The MUs are used as the basic features for the classification scheme described next.

Learning the "structure" of the facial features

The use of Bayesian networks as the classifier for recognising facial expressions was suggested by Chen et al. [42] and [50], who used Naive Bayes (NB) classifiers and who recognised the facial expressions from the same MUs. When modelling the described facial motion features, it is very probable that the conditional independence assumption of the Naive Bayes classifier is incorrect. As such, learning the dependencies among the facial motion units could potentially improve classification performance, and could provide insights as to the "structure" of the face, in terms of strong or weak dependencies between the different regions of the face, when subjects display facial expressions.

In our approach, instead of trying to estimate the best a-posteriori probability, we try to find the structure that minimises the probability of classification error directly. The basic idea of this approach is that, since we are interested in finding a structure that performs well as a classifier, it would be natural to design an algorithm that uses classification error as the guide for structure learning. Consequently, we further leveraged on two properties of semi-supervised learning: (1) the unlabeled data can indicate incorrect structure through degradation of classification performance, and (2) the classification performance improves with the correct structure. Thus, a structure with higher classification accuracy over another structure indicates an improvement towards finding the optimal classifier. The details of our analysis were presented in [51] and here we only briefly review the important issues that support understanding the classification component of our system.

To learn the structure using classification error, we adopted a strategy of searching through the space of all structures in an efficient manner while avoiding local maxima. As there is no simple closed-form expression that relates structure with classification error, it is difficult to design a gradient descent algorithm or a similar iterative method. Even if we did that, a gradient search

algorithm would likely find a local minimum because of the size of the search space. The solution followed in our system is the stochastic structure search (SSS) algorithm [51].

First it is necessary to define a measure over the space of structures which we want to maximise:

Definition The *inverse error measure* for structure S' is

$$inv_e(S') = \frac{\frac{1}{p_{S'}(\hat{c}(X)\neq C)}}{\sum_S \frac{1}{p_S(\hat{c}(X)\neq C)}}, \tag{2.1}$$

where the summation is over the space of possible structures, X represents the MU's vector, C is the class space, $\hat{c}(X)$ represents the estimated class for the vector X, and $p_S(\hat{c}(\mathbf{X}) \neq C)$ is the probability of error of the best classifier learned with structure S.

We used Metropolis-Hastings sampling to generate samples from the inverse error measure, without having to ever compute it for all possible structures. For constructing the Metropolis-Hastings sampling, we defined a neighbourhood of a structure as the set of directed acyclic graphs to which we can transit in the next step. Transition is done using a predefined set of possible changes to the structure; at each transition a change consists of a single edge addition, removal, or reversal. We defined the acceptance probability of a candidate structure, S^{new}, to replace a previous structure, S^t as follows:

$$\min\left(1, \left(\frac{inv_e(S^{new})}{inv_e(S^t)}\right)^{1/T} \frac{q(S^t|S^{new})}{q(S^{new}|S^t)}\right) = \min\left(1, \left(\frac{p_{S^t}}{p_{S^{new}}}\right)^{1/T} \frac{N_t}{N_{new}}\right) \tag{2.2}$$

where $q(S'|S)$ is the transition probability from S to S' and N_t and N_{new} are the sizes of the neighbourhoods of S^t and S^{new}, respectively; this choice corresponds to equal probability of transition to each member in the neighbourhood of a structure. This choice of neighbourhood and transition probability creates

a Markov chain which is aperiodic and irreducible, thus satisfying the Markov chain Monte Carlo (MCMC) conditions [180]. T is used as a temperature factor in the acceptance probability.

Roughly speaking, T close to 1 allows acceptance of more structures with higher probability of error than previous structures while T close to 0 mostly allows acceptance of structures that improve probability of error. Additionally, a fixed T amounts to changing the distribution being sampled by the MCMC, while a decreasing T is a simulated annealing run, aimed at finding the maximum of the inverse error measures. The rate of decrease of the temperature determines the rate of convergence. Asymptotically in the number of data, a logarithmic decrease of T guarantees convergence to a global maximum with probability that tends to one.

The SSS algorithm, with a logarithmic cooling schedule T, finds a structure that is close to minimum probability of error. We estimate the classification error of a given structure using the labelled training data. Therefore, to avoid overfitting, we added a multiplicative penalty term derived from the Vapnik-Chervonenkis (VC) bound on the empirical classification error. This penalty term penalises complex classifiers thus keeping the balance between bias and variance (for more details we refer the reader to [51]).

Please note that we decided to use this particular tracker due to its proven robustness and its ability to cope with non-frontal faces (up to 30% in head pose change). There were several other alternatives, mostly based on AAM (see for example [269] or [43]) but these systems require training and have difficulties in coping with the situations that were not present in the training set.

2.1.3 Analysis

This section presents the analysis of facial activity for detecting personal highlights of video contents.

Figure 2.3: Recording facial expressions of a viewer (Left) watching a video clip (Right).

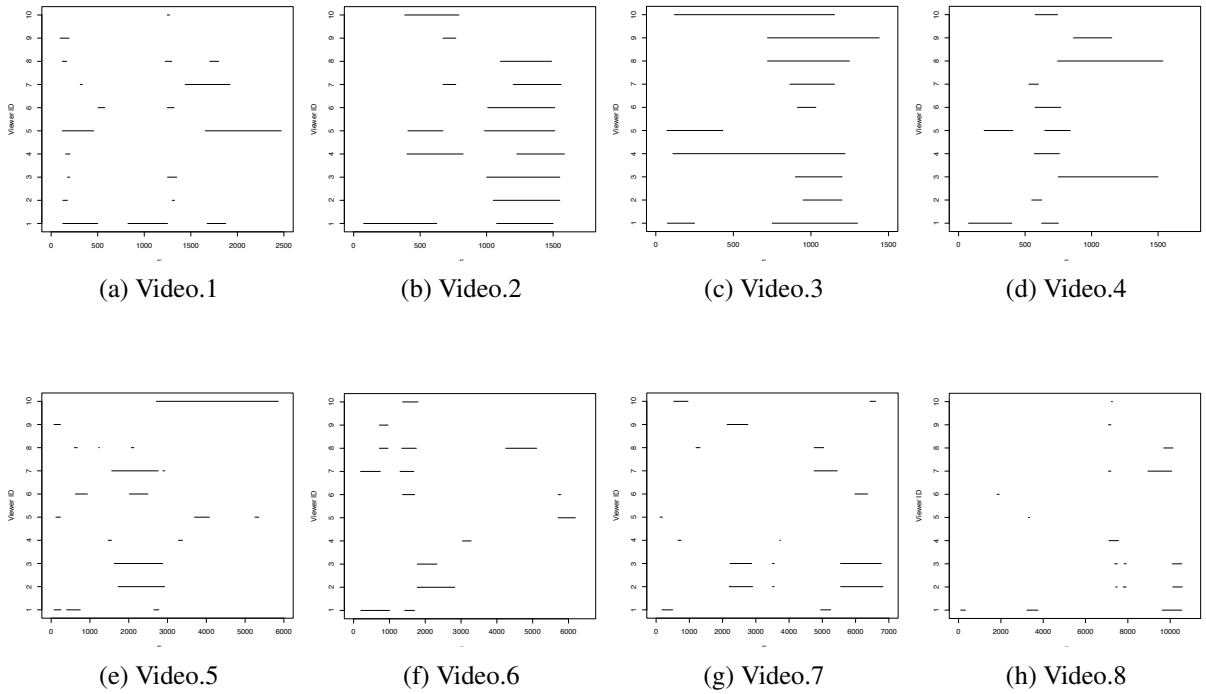
Participants and video clips

Ten people, all employees in the same software development company (holding different positions) agreed to participate in the experiment. Out of the ten, five were female and five were male. All participants were between the ages of 24 and 43, and were free from any obvious physical or sensory impairment. We used eight video clips taken from the contents in different genres. The code, duration, and brief description of the video clips, are given in Table 2.1. All videos had 25 frames per second.

The recording of facial activity was carried out in a room where a conventional video camera was set on top of a TV set. It should be noted that all video clips were new to the participants. The content video and the recording of facial activity were synchronised for subsequent analysis (See Figure 2.3). The facial activity videos were exported to 360x240 pixels AVI format with 25 frames per second (same as the content video clips).

Table 2.1: Description of video clips

Code	Length	Description
Video.1	01:43.5	Promotion Video of a pop song. Most parts are slow scenes where a singer is walking downtown while singing. There is a colour effect on the picture which tones the colours to green and yellow.
Video.2	01:20.0	Documentary of a man with physical impairment demonstrating day-to-day activities. Calm background music with no speech. Visually similar across the clip. A short subtitle at the beginning introducing the contents.
Video.3	01:36.4	Documentary of people with physical impairment. Scenes of dancing with a wheelchair (First half) and travelling to the river (Last half). Calm background music with no speech (Similar to Video.2). A short subtitle at the beginning introducing the contents.
Video.4	00:39.0	Comical TV commercial of a beer. Night scenes and inside scenes with background noise of insects. Speech from three people and narrator at the end. No music. Two scenes were interwoven.
Video.5	04:29.2	A car chase scene from an action film. Upbeat background music with many sound effects of siren, scratching tires, crash, etc. Speech from four people. Many fast moving short shots.
Video.6	04:48.2	Scenes from a comedy drama film. Two scenes were interwoven: a talkshow with one presenter, five guests on the stage, and large audience; and a scene introducing the background of the main character. Mainly speech with many short shots.
Video.7	04:43.4	An action scene at night from a Sci-Fi film. Two groups of people are shooting and fighting. Many sound effects (guns, helicopter, breaking glasses, etc.) but no background music. Some shouts and screams in fast moving shots.
Video.8	07:03.6	Scenes from a soap drama. Amateur football game scene (60%), many conversations between people (30%), driving a car (10%), etc. No background music, but noise from the audience in the football game scene.



(a) Video.1 (b) Video.2 (c) Video.3 (d) Video.4

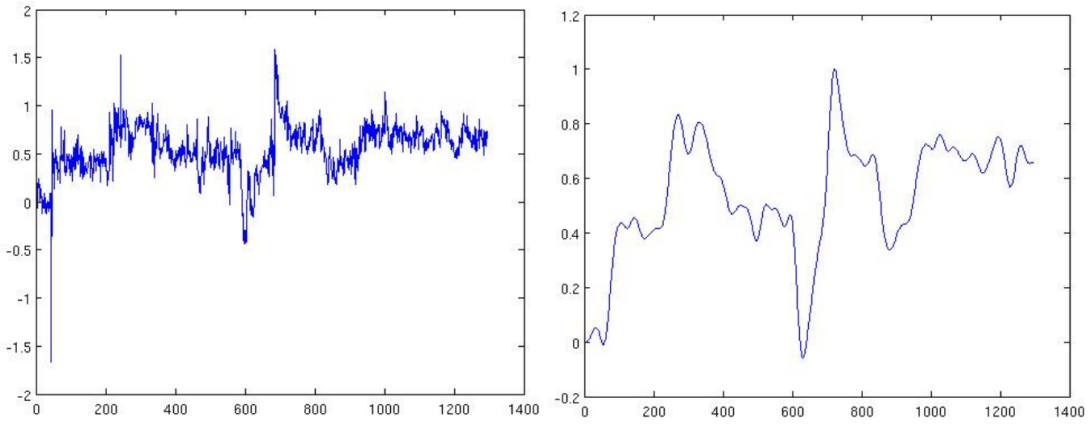
(e) Video.5 (f) Video.6 (g) Video.7 (h) Video.8

Figure 2.4: Annotation of personal highlights (Video.1 to Video.8)

Highlight annotations

We obtained the manual annotations of highlight scenes from participants to evaluate the effectiveness of facial motion units. After the end of a video clip, participants were presented with a simple video annotation tool where they could select parts of video clips. Participants were allowed to annotate as many separate scenes as they found it necessary as highlights. The results of the manual annotation can be found in Figure 2.4, where the X-axis represents the frame number of video clips and Y-axis represented the viewer ID. Note that the frame length denoted by the X-axis varies across the video clips.

As can be seen, there was a high level of consensus as to where a highlight is present in Video.2. As summarised in Table 2.1, Video.2 (shown in Figure 2.3) was a documentary of people with physical impairment. In the frames



(a) Original Data (b) Kaiser Window

Figure 2.5: Effect of Kaiser Window on MU1 feature.

between 1000 and 1500, one of the people skillfully folded a piece of paper using their feet. Most viewers selected this scene as the highlight of the video clip. However, such consensus did not appear to be common in most of the rest of videos. This observation is important since this suggests that people can find different parts of videos as the highlight, which is the major assumption made in this paper.

Facial features

We analysed a total of 20 facial features in this study. They included 12 motion units (denoted as MU1 to MU12), a combination of the 12 vectors (denoted as MU1-12), and 7 facial expression categories (Scared, Angry, Disgusted, Happy, Neutral, Sad, and Surprised). For each of the facial activity videos, a vector value of motion or probability of emotion categories were produced by the methods described in Section 2.1.2. We then applied a Kaiser Window process on the outputs of facial features in a similar fashion to [112]. The effect of smoothing on the original data can be found in Figure 2.5.

Our hypothesis was that an effective facial feature should produce a large

motion or high probability of emotion category to detect personal highlights of videos. Therefore, we see this as a ranking problem where the video frames are ordered by the vector value or category probability. Consequently, we used a scoring function called Average Precision [122, 290] to measure the effectiveness of facial features for personal highlight detection. Average Precision, $AvgP$, is one of the major performance measures in the field of Information Retrieval, and it is calculated in the following manner:

$$AvgP = \frac{\sum_{r=1}^{N} P(r)}{H} \tag{2.3}$$

$$P(r) = \frac{h(r)}{r} \tag{2.4}$$

where r is the ranking position of a frame, N is the ranking position of the lowest ranked highlight frame, $h(r)$ is the total number of highlight frame found up to the rank r, $P(r)$ is the precision at the rank r, and H is the total number of highlight frames annotated by individual participants.

2.1.4 Results and Discussion

This section reports the results of the analysis and discusses the implications of our findings on the design of personal highlights detection technologies for video contents.

Facial activity

The first analysis carried out was the performance of the 12 motion vectors to detect personal highlight scenes in video clips. The result is shown in Table 2.2. Motion unit IDs (MU ID) are based on the numbers shown in Figure 2.2. The values in the table are the mean of average precision of all video clips. The bottom row of the table shows the performance of a feature which combined the magnitude of 12 motion vectors. To highlight the effect of facial parts, the

features are divided into four parts: Mouth, Cheeks, Eyes, and all. We consider the mouth as the lower part of human face and the cheeks and eyes as the upper part of the face.

Table 2.2: Mean Average Precision of motion vectors. Those highlighted in bold are the best performance in individual viewers.

Facial Part	MU ID	Viewer 1	2	3	4	5	6	7	8	9	10
Mouth	1	.220	.098	.166	.116	.150	.058	.097	.186	.103	.121
	2	.278	.104	.172	.220	.135	.096	.083	**.233**	.073	.111
	3	.325	.113	.171	.098	.171	.059	.138	.162	.135	.122
	4	.192	.078	.208	.075	.088	.059	.139	.103	.118	.090
	5	.175	.072	.127	.099	.095	.050	.097	.149	.055	**.131**
	6	.187	.134	**.225**	.108	.097	.068	.103	.122	.130	.109
Cheeks	9	.195	.097	.197	.148	.121	.092	.079	.145	.087	.075
	10	.325	.139	.150	**.223**	.264	.065	.094	.176	.059	.093
Eyes	7	.147	.090	.129	.091	.097	**.143**	.132	.198	**.200**	.118
	8	**.404**	.104	.130	.163	.251	.072	.088	.145	.071	.089
	11	.316	.101	.155	.077	**.337**	.094	.104	.169	.172	.078
	12	.302	**.148**	.145	.078	.207	.066	**.163**	.145	.177	.096
All	1-12	.240	.127	.123	.078	.090	.051	.102	.135	.052	.095

There are several observations from the result. First of all, the most useful features to detect personal highlights significantly vary across the viewers. This suggests that people's facial activity to react to their highlight scenes can be indeed very different. Second, relatively speaking, the motion units in the upper part of human face appear to be more indicative of personal highlights than the lower part. Although the best performing features varied across viewers, seven out of ten were based on the upper part of human face, which included eyes (MU7, 8, 11, and 12) and cheeks (MU9 and 10). Of those, the MUs around the eyes had the largest number of best performing cases. This suggests that the effectiveness of motion units across the 12 points are not equal, and a greater level of attention to the upper part of human face might allow us to capture

individual preferences. Finally, the performance of `MU1-12` suggests that a simple addition of all motion vectors was not sufficient for accurate estimation of personal highlights.

Table 2.3: Mean Average Precision of emotion categories. Those highlighted in bold are the best performance in individual viewers.

Emotion	Viewer									
ID	1	2	3	4	5	6	7	8	9	10
Afraid	.228	.094	.137	.077	.145	.046	.078	**.211**	.062	.086
Angry	.225	.149	.187	.071	.174	.051	.119	.144	.049	.089
Disgusted	**.336**	.144	**.264**	.077	.147	.042	**.127**	.127	**.195**	.091
Happy	.238	.101	.208	.086	.099	.050	.107	.170	.041	.069
Neutral	.256	**.224**	.233	.104	**.234**	**.183**	.086	.152	.143	**.307**
Sad	.296	.134	.210	.100	.122	.044	.107	.161	.077	.182
Surprised	.258	.138	.126	**.110**	.179	.049	.069	.160	.078	.061
Best MV	.404	.148	.225	.223	.337	.143	.163	.233	.200	.131

Comparison to facial expression features

The second analysis compared the performance of motion vectors to that of emotion categories. The results are shown in Table 2.3. In the bottom row of the table are the best performing MUs from Table 2.2 for reference.

Unlike the performance of motion vectors, most of the best performing features in the emotion categories were based on the `Disgusted` and `Neutral` categories. However, if we compare these performance to the best MU features, we can observe that it was the Neutral feature which often outperformed the motion vector features. We speculate that the performance of Neutral category is partly due to the fact that many frames are categorised as Neutral when no particular facial activity was detected. Therefore, the Neutral category was more likely to perform better than other categories.

Overall, the comparison to the emotion categories suggests that some users can be modelled by a single point (motion unit) while others need multiple points (i.e., emotion category) to model their affective states.

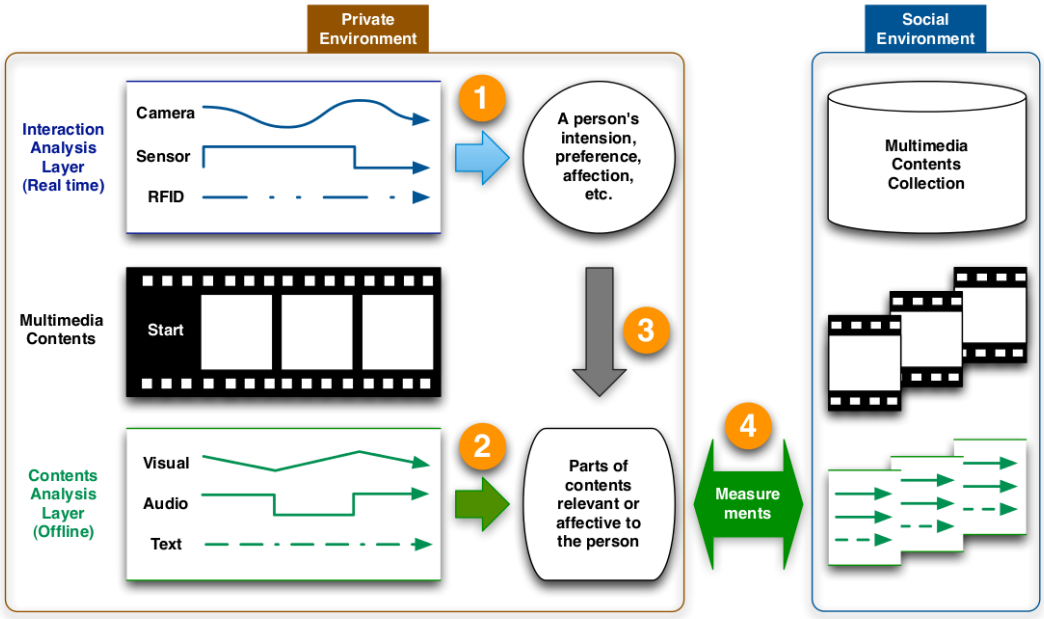


Figure 2.6: A framework for the multimodal approach to multimedia personalisation

On scalability

We have looked at people's facial activity to detect the personal highlights in video clips. This can be seen as a subtask of affective video summarisation based on human-centred multimodal approach [130]. A limitation of multimodal approach which exploits physiological aspects of human beings using various sensors is the scalability. Unlike the content analysis approach, we can only collect the data while the users engage with multimedia contents. While our approach was using only a conventional webcam which is much less obtrusive than other approaches, the limitation still applies. In our previous work, this issue was briefly discussed as follows (Note that FX stands for facial expression in the following quote):

> "we need to explore ways to leverage user based information in a practical fashion. One way might be the combination with content based approaches. For example, the highlight scenes are determined by FX based models in unobtrusive way, but the scenes were represented by low level feature models. This will allow us to generate a personalised summary for unseen videos by measure the similarity between existing FX profile and new video contents." [136]

This section expands our view of this issue by looking at the multimodal interaction analysis of multimedia contents in a larger context, which is illustrated as a research framework in Figure 2.6.

The framework assumes two major environmental parts, namely, a private environment and social environment. The collection and storage of an end-user's multimodal information should be carried out in the private environment given that many of the recordings can contain sensitive data. On the other hand, the majority of multimedia contents is available online as the social environment. A key issue is to bridge these two environments in a practical way.

The framework broadly divided the analysis into two layers. One is the interaction analysis layer which includes the multimodal interaction analysis presented in this paper, or others with various sensory devices as described in Section 2.1.1. The analysis of this layer tends to be carried out in realtime when the end-user engages with multimedia contents to capture a user's affection, preference, intention, and other user profile information. Another is the content analysis layer which includes the analysis of visual, audio, and textual data extracted from multimedia contents. This layer's analysis can be done in offline to capture the characteristics of multimedia contents at various levels.

Given that the interaction analysis layer can provide a rich representation of user profile information, one way to scale the multimodal interaction approach is to map the significant parts (e.g., affective) of multimedia contents onto the representation of the content-analysis layer. Once this mapping is successfully carried out, then a measurement such as similarity measure can be done with all the other contents available in the social environment at the content-analysis layer. We do not claim that this is the only way to further our research. For example, a successful mapping of the interaction analysis layer to the content-analysis layer can be challenging. However, it is clear from the framework that there is ample room for further investigation to achieve a scalable multimodal approach to personal highlight detection and affective multimedia summarisation.

2.1.5 Conclusions and Future Work

Detecting a personal highlight of multimedia contents is a key research issue for affective multimedia analysis and summarisation. We proposed a facial activity-based approach to personal highlight detection, which required only a conventional webcam system unlike other approaches. The preliminary analysis of our approach suggested that the motion vectors in a upper part of human face were more likely to be indicative of personal highlights than the lower part

of the face. We plan to develop a more sophisticated technique to detect personal highlights based on this finding in the future. We are also interested in the issue of mapping interaction analysis data to content analysis data to achieve a scalable multimodal profiling for multimedia contents.

2.2 From Facial Expressions to UX evaluation

In the last decade there has been a widespread move in HCI to consider emotional aspects of User eXperiences (UX) alongside the standard usability requirements [114]. This move has brought forward a need for new instruments to measure emotional responses to technology. Psychologists have long striven to overcome the difficulties of operationalising and measuring emotions, yet the HCI context introduces new complex challenges. Self-report instruments [26, 69, 128] are in need of a serious validation effort, and invasive physiological instruments contrast with the requirement of ecological validity of the evaluation settings. The measurement is further complicated by the low intensity emotional reactions often elicited in HCI settings [26, 69]. These reactions tend to be of a mixed nature [69] and are normally not accompanied by visually observable changes in a person state [68]. As such, they are difficult to be described using the basic emotion taxonomy [77] implemented in current tools for usability evaluation [67].

Furthermore, HCI researchers and practitioners are interested in emotions as a means to understanding dynamic interactions, whereas the bulk of research in psychology and marketing has considered static stimuli [69]. Finally, most HCI practitioners are likely to miss the theoretical and methodological background necessary to interpret self-reports or to operate complex and expensive physiological instruments.

This paper presents UX_Mate (UX Motion Activation Tracking Engine), a tool for the automatic detection of the dynamic user emotional state during inter-

action with technology. The system fulfills many requirements of UX research:

1. it does not need invasive devices nor controlled illumination settings;
2. it can be installed in any device featuring a commercial in-built video camera;
3. it tracks minute changes in the facial muscle activity of the user facilitating discrimination of mixed emotions, such as frustration or confusion;
4. it is cheap and does not assume heavy background knowledge.

We present two independent evaluation studies used to validate the performance of UX_Mate against that of skilled user researchers. The work focuses on usability evaluation of different interactive devices through facial cues; the approach can be extended to cover mixed feelings such as frustration, a feeling linked to interaction difficulties as in our scenario, flow and fun.

The main contributions of this research are:

- the development and evaluation of UX_Mate;
- a corpus provided to the community of synchronized and annotated videos of interactive behaviour and facial expressions, which can be used to ground research on the relationship between behaviour and emotion in HCI.

2.2.1 State of the Art

A large corpus of research has explored the computational implications of technology that *relates to, arises from, or deliberately influences emotions* [223]. However, less emphasis has been devoted so far to understanding how measures of emotions can support the evaluation of interactive devices, and to the validation of new measurement tools.

In this section, we summarise the state of the art of different approaches to emotion appraisal in UX evaluation.

Questionnaire Measures Dozens of affective inventories are available in the psychological literature. Questionnaires share the benefit of being ecologically valid, as they do not need to be administered in controlled settings. However, they can only provide a summary evaluation of past events and cannot capture the dynamics of the interaction. Due to the dissipative nature of emotion, this evaluation is likely to be affected by response bias.

One of the most extensively used questionnaires is PAD [191], which measures emotions on three independent dimensions (Pleasure, Arousal, Dominance) by means of a semantic differential scale. Although PAD is reliable, there are a number of difficulties associated with it. Firstly, it requires the respondents to provide 18 different ratings for each stimulus. Secondly, it requires statistical skills from the evaluators. Finally, the cultural frame of the respondent can bias verbal ratings: even small differences in wording can increase the level of cognitive noise and alter response patterns.

To alleviate these problems, HCI research has recently focused on shorter, non-verbal measurement tools. The ones most commonly used in evaluations [30, 47, 271] rely on visual representations of emotions. Examples are the Self-Assessment Manikin [30] and PrEmo [69]. Some research has also investigated the communication of emotion through tactile experiences with physical stimuli [128].

Yet, questionnaires may still have several issues, as many dimensions of user experience are not stable, singular judgments, but rather vary over the time course of the interaction.

Psycho-physiological measures A number of psycho-physiological measures, such as changes in blood volume pressure, skin conductance, heart beat rate, brain activity, and muscular activity responsible for changes in facial expressions, eye movement, or vocal tones, can be measured through various devices (e.g. sensors, electrodes, diodes). Facial expressions are a rich source of information

through which people convey emotion.

Research in psychology demonstrated that facial expressions show reliable correlation with self-reported emotions [146] and with physiological measures of emotion [62]. The most common approach used to measure patterns of facial movements in HCI relies on the detection of muscle activity through electromyography (EMG) [32, 116, 181, 234, 303]. Such studies investigate the electrical activity of several muscles (*corrugator*, *frontalis*, *orbicularis* and *zygomatic*) in a range of interaction tasks. Results are preliminary and at times contradictory, but overall they suggest a relationship between the activity of the corrugator (eyebrow movement) and zygomatic (mouth corner movement) with interaction events.

Overall, facial EMG was showed to be an effective method for tracking emotional changes over time. Yet EMG is not the expected panacea to the measurement requirements of HCI as it tends to provide exclusively information on emotional valence and does not provide clear information on the specific emotions elicited. Furthermore, there are still issues of external validity: facial expressions and self-reports do not always correlate [181, 316].

While physiological approaches share the benefit of being able to accurately capture changes in emotional states that cannot be measured using other methods [242], they all require specific expertise as well as special and expensive equipment [177]. To overcome these limitations, researchers started investigating how usability can be assessed by means of automatic analysis of facial expressions collected by video signal processing.

A commercial system is FaceReader, developed by VicarVision and Noldus Information Technology [67]. This tool, based on Ekman and Friesen's theory [77] of the Facial Action Coding System (FACS), can recognize six basic emotions (i.e., happiness, sadness, anger, disgust, surprise, and fear), which are returned as output of the video analysis. The system was tested in a usability evaluation triangulating data from three sources: questionnaires, human

judgments, and data from FaceReader [316]. The results showed consistency between FaceReader's output and expert-human judgment, while questionnaire data were not consistent with the other sources of emotional information. This lack of correlation can be due to the direct use of basic emotions, which are unlikely to be elicited in the HCI context. Moreover, FaceReader has a number of constraints related to illumination or background clutter, which can affect the output [100].

Although the first results obtained using video analysis are encouraging, further research is needed to face current limitations, with a particular concern about finding ways of exploiting psycho-physiological measurements in a cheap, non-invasive, and ecological fashion.

2.2.2 UX_Mate

UX_Mate (UX Motion Activation Tracking Engine), is a software tool developed for automatic assessment of UX by means of facial motion tracking. UX_Mate brings together the advantages of EMG and approaches based on video analysis since it does not require invasive devices and can be used in natural settings, including situations with critical or varying illumination conditions. Moreover, it exploits fine-grained facial motion tracking instead of relying on a fixed emotion classifier. This feature allows to take advantage of low-intensity, mixed emotions as the ones elicited in HCI.

UX_Mate does not focus on emotion recognition: it rather exploits global and local facial motion patterns building on a framework based on the anatomical analysis of the human face derived from FACS [77]. Since it was first proposed in 1978, FACS has been established as the most widely accepted coding system for facial expressions in a number of different research contexts and has been widely employed in vision based automatic analysis of human faces [67, 215, 240, 289]. FACS provides definitions for over 40 Action Units (AU), that correspond to the contraction or relaxation of one or more muscles

of the face and are responsible for facial appearance changes.

The subtle motion of facial muscles corresponding to fast transitory motion of AUs is a powerful indicator of micro-expressions [80], i.e., the involuntary expressions appearing for periods of time as short as 1/25 of a second. A distinctive property of such micro-expressions is that they can hardly be faked [80].

Despite this large success, FACS presents some limitations: human observers require specific training in order to exploit it [80] and it is very time consuming: coding 1 hour of video data requires 4 hours of work [35]. UX_Mate overcomes this limitation by a tracking system able to run in real-time. As opposed to other approaches, the system is robust to illumination changes.

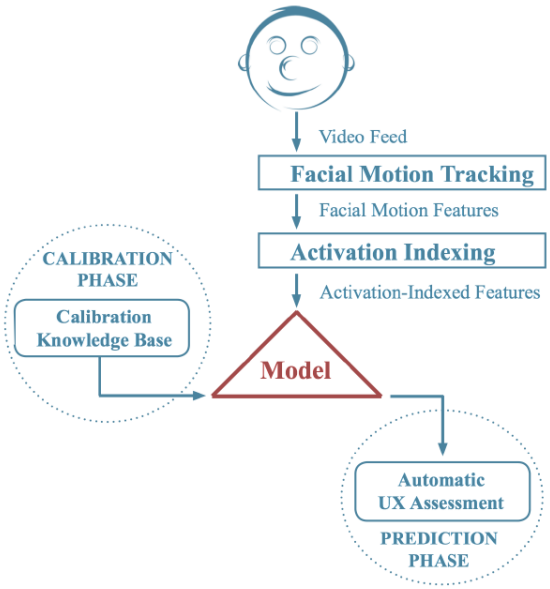


Figure 2.7: UX_Mate system overview.

A graphical overview of UX_Mate is portrayed in Figure 2.7. The video serves as input to the facial motion tracking module, whose output is then post-processed by the Activation Indexing module. The data generated by the Activation Indexing module serves as input to a generic Machine Learning module: in the calibration phase the activation-indexed features are used, in conjunction with the knowledge of the calibration task they refer to, to build a model of the

MU(s)	**Facial Movement**	***Activated Muscles***	**AU(s)**
1	Upper Lip Raise	*Levator Labii Superioriis*	10
2	Lower Lip Raise	*Depressor Labii Inferioris*	16
3,5	Right/Left Mouth Corner Horizontal Deformation	*Risorius*	20
4,6	Right/Left Mouth Corner Vertical Deformation	*Zygomatic Major, Levator Anguli Oris*	12,13
7,8	Right/Left Eyebrow Deformation	*Frontalis, Corrugator, Depressor Glabellae, Depressor Supercilii*	1,2,4
9,10	Right/Left Cheek Deformation	*Orbicularis Oculi (Pars Orbitalis)*	6
11,12	Right/Left Lid Deformation	*Levator Palebrae Superio Glabellaeris, Orbicularis Oculi (Pars Palpebralis)*	5,7

Table 2.4: Description of Motion Units

user's reactions. Such model is then used, in the prediction phase, to automatically assess UX in the tasks under evaluation. Manually labeled data is used in this paper to evaluate UX_Mate's performance.

Facial Motion Tracking System

The facial motion tracking system endorses a framework inspired by FACS: 12 Motion Units (MU) are defined in correspondence to one or more Action Units. The tracking information refers to the movement of these 12 MUs, corresponding to a subset of AUs defined in Table 2.4. This subset has proven to be sufficient for automatic facial expression recognition [49].

The key difference between AUs and MUs is that the latter not only represent activation of the unit(s), but also magnitude and direction of motion, making the measurement more informative. The tracking implementation we employ in UX_Mate is an adapted version of the algorithm proposed in [274]. This algorithm makes use of a Piecewise Bezier Volume Deformation (PVBD) model in a fixed camera environment; video data is captured by webcam. In the initial-

ization stage, a near-frontal and neutral face is captured and a 3D facial mesh model is fitted on the face. Such a model consists of sixteen surface patches (which are guaranteed to be continuous and smooth) embedded in Bezier volumes. The control points of the surface patches are the facial points of interest represented by the MUs.

On the chosen initialization frame, the motion vectors are set to zero; on subsequent frames, a template matching method is used to estimate the two-dimensional motion of the mesh nodes of interest. The mesh is then updated by projecting the two-dimensional motion information onto the three-dimensional face model. For each processed video frame, the facial motion tracking module outputs a 12-dimensional vector. The values in the output vector correspond to the direction and magnitude of motion for the corresponding MUs. We will refer to such values with the term features from now on.

For the sake of the current studies, we computed two variables based on the combination of different MUs, namely, confusion and frustration. The algorithm was based on FACS based research [77] showing a correlation between AU12 (lip corner puller) and frustration, and between the combination of AU4, AU7 and AU12 and confusion. Frustration features were computed for both mouth corners by calculating the length of the vector resulting from the addition of the respective horizontal (MU3/MU5) and vertical vectors (MU4/MU6). Confusion was computed by the quadratic mean of the individual MUs (eyebrows, eyelids, and mouth corners). Additionally, a measure of the overall facial activity was computed by adding all motion units values.

New variables can be added to cover a larger set of feelings linked to interaction events of interest. The described tracking algorithm has been previously used in several application scenarios, e.g. for affective video summarization [137].

Activation Indexing

The purpose of the Activation Indexing module is to detect MUs' activation at each frame, and subsequently score each task in terms of activation levels. The module computes the mean and standard deviation for each motion unit. The Activation Index for a MU is given by the count of frames (1/25 of a second) where its absolute value goes above one standard deviation over the mean. Such a computational approach is based on the procedure suggested in [116]. The criterion of one standard deviation is justified by standard in psychological testing, where these values are considered to be out of the normal range.

Machine Learning Module

The activation-indexed features are used as input data for a generic machine-learning module written in the Java programming language and based on the WEKA [107] data mining tool.

The modular approach adopted in designing UX_Mate makes it possible to connect data output to any machine learning toolkit with a Java interface. This module returns a prediction on the level of occurrence of a given event. A high level of flexibility is achieved through the use of calibration tasks, i.e. short sessions carefully designed in order to elicit specific reactions according to the goal of the evaluation to carry out. In the example described in this paper, the participants were tested with two short tasks designed to elicit a variable level of difficulty, but the tasks can easily be updated to fit different requirements.

2.2.3 Pilot Study

A pilot of UX_Mate was run during the evaluation of four Media-Players. The videos of participants faces served as input for UX_Mate, which returned feature values representing the level of activation of the motion units and the compound variables tracked. These values were used as predictors of performance

(interaction time and errors). Furthermore, a measure of external validity was collected asking three human observers to judge the difficulty of the tasks based on the videos of the users' faces.

Method

Participants Fifteen Masters students (14 M; mean age = 26.4 years) of a local University were involved, on a voluntary basis, as participants in the study. All of them reported at least three years of experience with different media players, but none had ever used the ones tested in the evaluation.

Procedure The study was conducted in several places, including rooms in the hall of residence or university offices using the participants' own computers. Four Media-Players were tested: iTunes, MusicBee, Songbird, and MediaMonkey. They each had a different look-and-feel, and level of usability and functionality. The media-players were installed on the participants' laptops alongside a program for synchronized video and audio recording of faces and screen actions.

Before the study, the participants signed a consent form stating that their face would be videotaped but with no reference to why. Then they performed three tasks on each media player: importing a folder to the library; finding a song and playing it; adjusting the equalization of a song. Media player order was counterbalanced across participants, while task order was kept constant. After task completion, participants filled in the UX questionnaire referring to the media player they had just used. At the end of the study, they chose one media player and committed to use it instead of their usual program for the following month.

During the experiment, the evaluator remained in the room with the participant in order to intervene in case of technical problems, but the amount of the interaction with the participant was kept to a minimum.

Measures The study collected three classes of measures: performance data, self-reports on user experience, and facial cues extracted by UX_Mate. Performance data (errors and time) were obtained from the expert analysis of the interaction videos.

Users' interaction path for each task was compared to an ideal task analysis describing optimal performance (i.e., the procedure allowing reaching the goal with the least number of actions). Several analyses were performed for each task and media players to account for different possibilities to achieve the same goal (e.g., direct manipulation vs. menu selection). All user actions not matching optimal performance were counted as errors.

The questionnaire was composed of three parts addressing UX evaluation, information about participants previous usage of Media-Players, and demographic data. Media players were evaluated for individual dimensions of UX and summary judgement. A definition of each dimensions and item wording is reported in Table 2. Dependent variables were computed averaging items of individual scales (all $\alpha > .80$).

Based on literature analysis, we selected 4 MUs (MU4/6 describing the movement of the zygomatic major and MU7/8 describing the movement of the corrugator) which could better describe the facial expressions of people facing difficulties. Furthermore, we analysed the compounded indexes describing frustration and confusion.

Results

Performance data A sample of 168 tasks was collected and used for performance evaluation. The average number of error per task was 3.98 (SD= 7.14), ranging from a minimum of 0 to a maximum of 39. The distribution shape of the variable error and time was improved by computing the square root of each data-point. The normalized variables were analyzed by two ANOVAs with media-player (4) and task (3) as between-subjects factor. Post-hoc comparisons

were based on the Least-Significance Difference method. Partial eta-squared (p2) was used as an estimate of effect size.

The ANOVA on error returned a significant effect of media-player ($F_{(3,156)}$ = 6.73, $p < .001$, η_p^2 = .12) and task ($F_{(2,156)}$ = 7.83, $p < .01$, η_p^2 = .09). The ANOVA on time returned a significant effect of media-player ($F_{(3,156)}$ = 5.85, $p < .001$, η_p^2 = .10) and task ($F_{(2,156)}$ = 7.70, $p < .001$, η_p^2 = .09). In both analyses, post-hoc tests indicated that MusicBee was significantly worse than all other systems (with no significant differences between them), and that task 2 was significantly easier than the other tasks.

Questionnaire data The scores of the 6 UX dimensions tested in the study were entered as dependent variables in 6 repeated-measures analysis of variances, with media-player (4) as within-subjects factor. The analyses returned a significant main effect of media-player for all variables. The F values ranged from 7.55 (p = .001, p2 = .35) for the functionality score to 12.75 ($p < .001$, η_p^2 = .48) for classical aesthetic.

iTunes and Media-monkey were preferred in all UX dimensions, with no significant differences between them. Songbird and MusicBee scored negatively, with significant differences favoring SongBird on usability, functionality, pleasure and summary judgement. This trend of results is consistent with participants final choice of the media-player to use for the next month. Ten people decided to use iTunes, nobody selected MusicBee.

Performance/Questionnaire correlation A correlation analysis was performed on total number of error and UX dimension scores. The correlation matrix reported only 2 small but significant negative correlations for functionality and classical aesthetics ($p < .05$). The analysis was repeated for each individual task, showing that the number of errors committed in task 1 and task 2 were not correlated with any UX dimension. The number of errors in task 3 showed 4 significant

correlations, for usability, classical aesthetics, functionality and pleasure.

Facial cues Three separate ANOVAs were performed to analyse the conjoint effect of the MUs measuring activity in the right or left hand-side of the face. In particular, we analysed MU4 and MU6 (eyebrow deformation), MU7 and MU8 (mouth corner movement), and the two frustration index as a function of media-layer (4) and task (2). The analysis on the MUs related to the zygomatic muscle (MU4 and MU6) returned a main effect of media-player ($F_{(6,254)}$ = 2.22, p ¡ .05, η_p^2 = .05). Post-hoc analysis indicated significantly more activation of both MUs during the interaction with MusicBee. The analysis on the MUs related to the corrugator (MU7 and MU8) returned no significant effects.

The analysis on frustration returned a multivariate main effect of media-player ($F_{(6,258)}$ = 3.04, $p < .01$, η_p^2 = .07). The difference was due to MusicBee, who elicited more frustration than all other systems. Identical results emerged from the Anova on confusion. The effect of media-player was significant ($F_{(6,258)}$ = 5.26, $p < .01$, η_p^2 = .11) and MusicBee was identified as the worse system. Correlations between the facial features and the number of errors are reported in Table 3. There was a strong correlation between eyebrows movements (corresponding to the corrugator muscle activity) and the number of errors. The correlation between the mouth and the errors was lower, but still significant.

A regression analysis was run to investigate the influence of the 4 motion units in predicting the number of errors. Using the enter method a significant model emerged ($F_{(4,140)} = 63.2p < .0010$). The model explained 65% of the variance. Table 4 gives information for the predictor variables entered in the model. MU8 was the strongest contributor to the model, MU4 and MU6 were also significant but contributed substantially less in the prediction.

Human Validation Three independent observers (2 M, 1 F; mean age: 27 years) were involved in an external validation of UX_Mate. The objective was to understand the level of agreement between UX_Mate and the human capability to detect task difficulty by observing the faces of users performing the tasks. The videos collected during the experiment were divided into six blocks of 24 clips each. The order of clips was randomized in a way that each block contained clips recorded by all the participants.

The three observers watched and rated all the six blocks of videos; the order of presentation of the blocks was randomized. For each video clip, the observers were asked to rate how difficult is the task the person shown in the video is dealing with; judgement were expressed using a 5-point Likert scale.

Only one judge performed well against the ground truth (i.e., actual numbers of errors in the interaction). Her scoring was highly correlated with the number of errors ($r = .48, p < .001$) and with the facial features as extracted by UX_Mate (r ranging from .19 to .33). This person was the only one who has had specific training on FACS. The other evaluators' scorings did not have any significant correlation with errors and with the output of UX_Mate.

Discussion

This pilot reports a preliminary evaluation of UX_Mate on the evaluation of 4 media players triangulating different measurement techniques. Overall, all measures (performance data, questionnaire scores and facial motion features) could consistently identify the worst system (i.e., MusicBee), although task variability was detected only by performance data and not by facial motion features.

We found evidence of correlations between facial expressions as detected by UX_Mate and the number of errors. In particular, the regression analysis suggested that movements of the eyebrows were a powerful predictor of error occurrence, whereas mouth expressions were weaker and could not discriminate the worst system if used in isolation. This effect may be due to the tendency of

participants of touching the lower part of their face, particularly in the moment of tension leading to video occlusion and hampering automatic detection of their emotional state.

The human-validation study suggested that UX_Mate can perform a task which is difficult and extremely expensive for human evaluators. The inference of behavioral cues based on facial cues requires a high level of specialization, as highlighted by the differential performance between trained and untrained observers. Overall, untrained observers appeared to express random evaluation with no correlation with each other, actual behaviour, or facial movements as highlighted by UX_Mate.

Overall, facial expressions showed a strong inter-individual variability within the sample. Some users had a very expressive and clear facial vocabulary, while other users had almost no apparent variations. This individual difference suggested the need of training UX_Mate to recognize the user facial expressivity before the evaluation task via a set of calibration tasks. Such tasks need to be designed carefully to identify the behaviour of interest, and need to be extremely short not to disturb the evaluation process.

In the validation study, we introduced two 10 second calibration tasks to gain knowledge about a participant's facial behavior in situations of different complexity. The study also highlighted the limitations of post-test questionnaires in assessing the dynamic of experience evolution over the time course of the interaction.

Indeed, the total number of errors and the subjective evaluations at the UX dimensions were only loosely correlated. On the contrary, we found evidence of a peak-end experience effect [142]: participants' evaluation tended to reflect their performance in the last task which was significantly more complex than the previous one. This result is consistent with a growing body of research, showing that when people construct summary judgements they are not only influenced by the average or the sum of their experiences, but that the final episodes and

their valence have a major influence on summary judgement [15].

2.2.4 Validation Study

A validation experiment was conducted on a set of videos collected during the evaluation of three social network services directed at people who study or work in universities or research institutions.

The objective of the study was to understand how well UX_Mate predicts task difficulty through automated analysis of video recordings of the face of the users. The performance of UX_Mate was calculated by comparing the system's prediction of error occurrence against expert-based annotation of the errors performed by participants while completing the tasks they were assigned. Participants' faces were recorded by the webcam of their laptops and the actions performed on the interface were captured. Webcam recordings served as input for UX_Mate and for training the system to predict usability issues on the basis of facial emotional cues. Screen captures were used for manual annotation of errors. Moreover, participant evaluation of tasks and websites was collected during the experiment by means of questionnaires and compared to system predictions and expert-based evaluation.

The experiment followed a similar procedure as the pilot study thus, only variations are reported.

Method

Participants Thirty-one students (26 M, mean age = 28.5 years) of a local University were involved on a voluntary basis as participants in the study. Almost all the participants declared to have experience in using social network sites except two (1M, 1F). None of them ever used the sites tested before the study.

Procedure The first step of the study consisted in the completion of a questionnaire assessing the participant's previous experience with academic social networks.

Then, UX_Mate was calibrated by means of a task in which participants were asked to retrieve information in two conditions: easy, the target information was presented within a well-formatted table; and difficult, the target information was hidden among irrelevant information and visual clutter [282]. In the easy condition, which was taken as baseline, minimum modifications of facial expression were expected, whereas in the difficult condition the nature of the task was expected to elicit changes in facial expression connected to task difficulty.

After calibration was completed, participants were provided with the links to the homepages of the three academic social networks (`academia.edu`, `researchgate.net` and `iamresearcher.com`) and asked to perform the following tasks on each of them: 1) create an account; 2) edit profile by uploading a photo; 3) search for a publication; and 4) delete the account. The websites were presented in a random order, while the order of tasks was the same for the three websites. After each task, the participants were asked to rate its difficulty on a 7-point Likert scale. After interaction with each website, users filled in the UX questionnaire.

Video analysis A detailed analysis was performed on the interaction video of each participant, annotating starting and ending point of every error. Following the operational definition adopted in the pilot study, all variations from optimal performance were scored as errors.

Double coding was performed on the entire sample and discrepancies were solved by discussion. A performance index was finally computed by dividing the number of video-frames spent in error and the number of total video-frames per task.

Results

Calibration task A set of t-test was run to compare the activation level of the 4 MUs correlated to task difficulty and of the compound variables between the two calibration conditions. Bonferroni corrections were applied.

All features analysed evinced a significant higher activation in the complex condition than in the easy one (Table 5).

Performance data A sample of 358 evaluation tasks was collected and analysed. The average number of errors per task was 0.86, ranging from a minimum of 0 to a maximum of 7. Some 49% of the tasks contained no error, 39% of them contained 1 or 2 errors, 8% contained 3 or 4 errors, and the remaining 3% contained 5 or more errors. On the average, participants spent 13% of the interaction time performing wrong actions.

The performance index was analyzed by an ANOVA with website (3) and tasks (4) as the within-subjects factors. The effect of task was significant ($F_{(3,75)} = 10.54, p < .001, \eta_p^2 = .3$). Post-hoc analysis showed that during task 1, participants spent less time performing wrong actions than during the other tasks. No other significant effects were returned.

Performance/questionnaire correlation All UX dimensions tested in the study reported high reliability values at the Cronbach test ($\alpha > .82$). Average scores were computed and used in the following analyses. The correlations between UX dimensions and perceived task difficulty indicated significant negative correlation only for one of the four tasks performed in the study (Table 6). For the other tasks, no associations were found.

The correlation matrix between perceived task difficulty and the performance index is reported in Table 7. The significant coefficients show a positive correlation between perceived difficulty and performance index only for half of the tasks analysed. The correlation between the percentage of task spent in error

(total and for the 4 tasks separately), and the 5 UX dimensions highlights the same peak-end experience effect as in the pilot study. Significant correlations appeared only with respect to the performance to the last task.

Automatic UX assessment

In order to assess the performance of UX_Mate, a bayesian model was trained using the Activation-Indexed data resulting from the calibration tasks and evaluated using the Activation-Indexed data resulting from the evaluation tasks. The model returned a set of predictions indicating if a task was simple or difficult. These results were compared against the expert-based annotations of the evaluation tasks. Tasks were marked as simple if they contained no errors, otherwise they were marked as difficult.

The confusion matrix, reported in Table 8, visualizes UX_Mate's performance in predicting simple or difficult tasks based on the presence of usability errors (annotated during the video analysis). The highest the values along the diagonal, the better the system's performance. It is evident that UX_Mate performed reasonably well: it correctly predicted the class of more than two thirds of the tasks.

In particular, it was able to correctly classify 101 tasks as simple, out of the 136 tasks marked as simple in the dataset (column 1). Conversely, it was capable to identify 151 tasks as difficult out of the 222 marked as difficult in the data set (column 2). Table 9 summarises the experimental results considering three standard performance indexes used in machine learning. Precision is the number of tasks identified correctly out of the total number of tasks. Recall is the number of correct results returned by the model, divided by the number of results that should have been returned.

The F1-Measure is defined as the harmonic mean of precision and recall, therefore its values provides an estimate of both variables. Overall, UX_Mate obtained a slightly higher precision in classifying simple tasks than in classify-

ing difficult ones. On the other hand, a much larger difference emerged with recall: 81% of the tasks marked as difficult in the dataset were correctly identified as such.

Discussion

The study provided encouraging evidence on the validity of UX_Mate as a tool to predict the occurrence of usability errors. Overall, the system was capable to discriminate between tasks containing and tasks non-containing errors in over two thirds of the cases. The study also confirmed the complexities of UX evaluation as assessed by questionnaire data.

Not only the summary judgment was strongly biased towards the performance of the final tasks, as previously highlighted in the pilot study, but also the pattern of association between different types of subjective data collected at different stages of the evaluation was complex. Significant correlations between perceived difficulty and UX dimensions were systematically found only for one specific task.

Finally, we found weak evidence of correlations between participants self-reports on task difficulty and occurrence of usability errors, showing that these two concepts may not always be related. We believe that the performance of UX_Mate can be improved by collecting a larger database of appropriate calibration tasks on which to train the model used by the machine learning module (Figure 2.7). It is fundamental that these calibration tasks are unambiguously linked to the interaction events extracted during the prediction phase. This aspect was a limitation of our study. Indeed, the calibration task we used, despite having the advantage of being extremely short and already tested in the literature [282], had only an indirect link with the event of interest implied by the equation: usability problems = increased difficulty. The mismatch was evident in the questionnaire results.

Further improvement to the performance of UX_Mate can be achieved by

setting more discriminative thresholds between the classes predicted by the machine learning module. The validation study reported in this paper addressed professionally designed websites, with less than 1 error per average task. As a consequence, we had to use two broad prediction classes, just considering presence versus absence of errors. Despite this sampling limitation, UX_Mate was still able to recognize the occurrence of interaction errors as tagged by expert evaluators even when the task was not perceived as difficult by the users.

2.2.5 Conclusions

This study presented UX_Mate, a modular system which tracks facial expressions of the users, interprets them based on pre-set rules, and generates predictions about the occurrence of target behaviour during HCI. Prediction is based on facial expression examples collected in the calibration phase. In this paper, we concentrated on prediction of errors occurrence from facial expressions linked to frustration and confusion.

In future research, we aim to extend this paradigm to other interaction feelings, such as enjoyment or boredom, in order to study what interaction features may be responsible for their occurrence. UX_Mate is designed in a modular way, allowing the evaluator to choose the machine learning algorithms, and the set of examples to train the system that best fit their particular needs. We have now collected a new large database of facial and interaction videos in an entertainment setting focused on feelings like flow, engagement and fun.

UX_Mate represents a preliminary yet important step towards the automatic assessment of User eXperience. We believe that the automatic assessment of facial expressions can be a powerful tool to support UX studies following a research paradigm based on the triangulation of different techniques, including human-based observation, self-reports, and facial expression tracking. In particular, UX_Mate can provide a cheap method to monitor the dynamic evolution of emotions in time, and counteract the tendency of questionnaires to anchor on

specific moments of the performance. An automatic tool can be an important help to untrained evaluators when they need to understand emotional reaction.

UX_Mate can perform a task which is difficult and extremely expensive for human evaluators. Indeed the pilot study demonstrated that inferring performance behaviour based on visual cues from the participant face is time consuming and requires a high level of specialization. Contrary to [316], we claim that UX research requires methods that can be applied beyond the laboratory settings of a usability laboratory. In this respect UX_Mate represents a unique and promising tool.

Chapter 3

Human Interaction in Small Groups: What and How Influence Whom

In [267] we presented a multimodal framework employing eye-gaze, head-pose and speech cues to explain observed social attention patterns in meeting scenarios. We first investigate a few hypotheses concerning social attention and characterize meetings and individuals based on ground-truth data. This is followed by replication of ground-truth results through automated estimation of eye-gaze, head-pose and speech activity for each participant. Experimental results show that combining eye-gaze and head-pose estimates decreases error in social attention estimation by over 26%.

Furthermore, in [256], we target the automatic recognition of personality states in a meeting scenario employing visual and acoustic features. The social psychology literature has coined the name personality state to refer to a specific behavioral episode wherein a person behaves as more or less introvert/extrovert, neurotic or open to experience, etc. Personality traits can then be reconstructed as density distributions over personality states. The personality states we are addressing are those corresponding to the Big Five traits. Different machine learning approaches were used to test the effectiveness of the selected features in modeling the dynamics of personality states.

This work is the very first computational approach to personality *states* recog-

nition.

3.1 Putting the Pieces Together: Multimodal Analysis of Social Attention in Meetings

Determining the direction of another person's attention is an important ability for humans. It not only provides salient information about the location of objects (food, predators), but also plays a fundamental role in many complex forms of social cognition such as visual perspective-taking, deception, empathy and the theory of mind [306], expression of intimacy and exercising of social control [151].

Gaze direction is an important cue for social attention and humans have evolved specialized neural mechanisms devoted to gaze processing [159]. However, it has been convincingly shown that there is more than just eye-gaze to visual attention; head and body orientation also significantly contribute towards deciphering another person's direction of attention [159]. While Perret *et al.* [220] developed an attentional model that integrates eye gaze, head and body directions in a hierarchical fashion, recent work [158] suggests these orientation cues are processed independently and combined so that one modulates the decision process concerning the others.

In this work, we investigate computational models of social attention by considering gaze direction, head orientation and speaking activity. We consider a number of hypotheses concerning social attention in meetings by analyzing results from ground-truth as well as automated analysis of four meeting videos from the 'Mission Survival II' corpus [185]. The following hypotheses are based on observations and presumptions stated in previous literature, but which have never been analyzed in great detail:

- H1: Attention is mostly given to the person sitting right in front of the observer. This hypothesis derives from an observation made in [261].

- H2: There exists a direct relationship between the verbal behavior of a person and the amount of attention he receives. This hypothesis derives from [216].
- H3: Use of eye-gaze in conjunction with head-pose improves accuracy of automated social attention estimation. This hypothesis directly derives from the above discussion.

We also attempt to characterize meetings and individuals based on the analysis of ground truth data. Finally, we describe automated methods to compute eye-gaze-cum-head-pose-based social attention, and replicate ground truth results by combining computed social attention estimates with speech data. To summarize, this is one of the first works to

1. Comprehensively analyze meeting videos by combining eye-gaze, head-pose and speech information. Past works have essentially focused on perfecting automated methods for computing visual social attention.
2. Automatically employ eye-gaze as a modality for estimating social attention using [285]. Previous works assume head-pose as the main indicator of social attention, mostly due to the difficulty in reliably computing eye-gaze. Experimental results show a significant increase in attention estimation accuracy when gaze cues are employed in conjunction with head-pose cues.

3.1.1 Related Work

Social attention has been extensively investigated under the rubric of focus of attention (FOA) in meetings [91]. Pioneering work is described in [260], where subjects' FOA is computed by combining head-pose information with *a-priori* knowledge about the number of participants and their relative positions. Assuming the head-pose to be the main indicator of a person's direction of attention, the algorithm employs a Hidden Markov model (HMM) to map FOA estimates

to real-world targets. This framework is extended to employ acoustic as well as visual cues in [261].

Prediction of focus of visual attention in dynamic meeting scenes is discussed in [294]. Shifts of FOA in spontaneous situations are studied for 10 videos with 35 possible attention targets. The most probable target is identified by mapping head-pose to its most likely gaze angle counterpart, to achieve 57% correct recognition of the visual target. Another approach to recognizing social attention in meetings from head-pose modeled using a Gaussian mixture model (GMM) as well as HMM, is discussed in [18]. FOA targets are not restricted to participants alone, but to environmental targets (*e.g.*, projector) as well, and results of saccadic eye motion modeling are exploited to model head-pose given the upper-body pose and effective gaze target.

An attempt to integrate contextual cues with head-pose information to compute FOA is presented in [19]. Instead of defining head-pose using pan and tilt angles with respect to the reference, the posterior probability density function of the different head poses is utilized to better fit the pose information to a given FOA target. For a corpus of 4 meetings of 15-27 minutes duration, each comprising 4 subjects, this algorithm improves performance by 5.4% over previous approaches.

Recent research has focused on automatic analysis of social aspects such as meeting roles, with specific emphasis on *dominance*, which characterizes a person's status within a group and the power he/she has within it. A study on the usefulness of non-verbal audio-visual cues when employed individually or in combination, for automated dominance estimation is described in [123]. Another work that discusses dominance estimation from meetings is [124], where the visual dominance ratio (VDR) measure is employed for automated dominance computation.

Brief analysis of related literature shows that (1) While past works have investigated and considerably improved on the usage of features such as head-

pose, an explicit analysis of social attention and multimodal cues to define **meeting characteristics** is missing. (2) Most works consider head pose as the main indicator of social attention, neglecting eye gaze primarily because of the difficulty in computing it. The next section describes the meeting videos used for analysis and the derivation of meeting characteristics upon analysis of the ground truth data.

3.1.2 'Mission Survival' Meeting Videos

We used data from the 'Mission Survival' corpus [185], a multimodal annotated collection of video and audio recordings in a lab setting. Each meeting consists of four participants seated around a table and engaged in the 'Mission survival' task, which is used in experimental and social psychology to elicit decision-making processes in small groups. The objective of the 'Mission Survival' task is to reach a consensus on how to survive a disaster scenario, *e.g.,* a plane crash in an uninhabited island. The group has to rank a number of (up to 15) items critical for survival, according to the participants. The consensus meeting scenario was chosen for the purpose of meeting dynamics analysis, which involves intensive engagement of the participants in order to reach an agreement, thus offering the possibility to observe a large set of social attitudes. All meetings are of 20-30 minutes duration, and recorded with four web cameras installed on the meeting table, while speech activity is recorded using close-talk microphones. Fig.3.1 shows an exemplar meeting scenario from the 'Mission Survival' data. Assuming that each participant directs his/her social attention targets included only the remaining three subjects, annotations were performed for the head-pose, eye-gaze and speech data to obtain the ground truth. Since the nature of the task involved choosing from a list of items, a 'self-attention' label, which denotes the state where a participant looks at the list provided to him/her, was also included in the annotation.

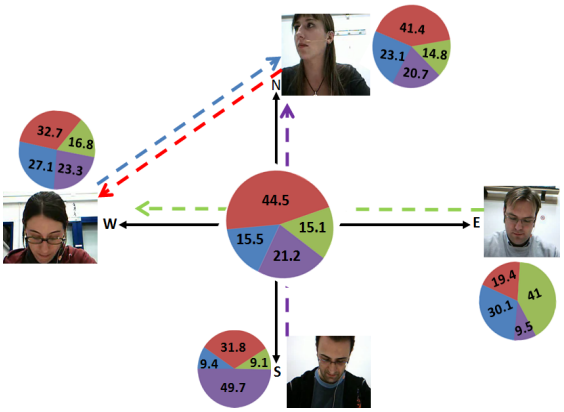


Figure 3.1: An exemplar meeting scene from the 'Mission Survival' dataset [185]. Color codes denoting subject locations are red (North), blue (West), violet (South) and green (East). The central pie-chart represents the distribution of speaking time, while pie-charts beside each participant denote the distribution of *attention given* by that subject to peers, including self-attention. Arrows denote direction of maximum *attention given* (excluding self-attention).

3.1.3 Inferences from Ground-Truth

Since eye-gaze is the most reliable social attention cue, we analyze eye-gaze and speech ground-truth data to derive inferences in this section. Fig.3.1 presents the distribution of social attention and speech activity for a meeting. Let A_i^j denote the *attention given* by subject i to j. Conversely, A_j^i denotes *attention received* by subject i from j. A_i^j and A_j^i may be expressed in minutes or as percentages. Henceforth, $i, j \in L, O, R$, where L, O and R denote the person located at the *left*, *opposite* and *right* respectively, with respect to the reference. Also, let $A^i = \sum_{\forall j, j \neq i} A_j^i$. denote the *overall attention received* by subject i. Likewise, $A_i = \sum_{\forall j, j \neq i} A_i^j$ denotes *overall attention given* by subject i to his peers.

Using only the eye-gaze social attention data, the first hypothesis we would like to validate is **H1:** ***The person opposite to the reference subject receives significantly more attention by her***. Indeed, in [261], experimental results suggest that there exists a considerable attention bias towards the person seated directly opposite, irrespective of who the speaker is. However, this aspect of social attention has not been studied in subsequent research works.

Next, we investigate the relationship between *attention received* and *speaking time*. If ST_i denotes the speaking time (expressed as percentages over the meeting duration) for the i^{th} participant, our second hypothesis is **H2: *The overall attention received by a participant is influenced by the amount of her speech activity*** or $A^i \; \alpha \; ST_i$, for a given subject. This hypothesis is based on the intuition that the persons who contribute most to the meeting should, in general, receive the more attention. The overall *attention received*, A^i, can therefore be divided into two components: $A^i_{(s)}$, which denotes the *attention received* when the subject is *speaking* and $A^i_{(l)}$, representing the *attention received* while *listening*.

Finally, we attempt to characterize and categorize meetings and persons based on the meeting statistics, namely, the mean and variance for the *speaking time* and *attention received*. Our assumption is that analyzing these measures we can characterize the type of meeting and how each person participates in the discussions.

Validation of H1

The distribution of A^j_i and A^i_j to subjects seated to the left, right and directly opposite is not biased as observed in [261], where the authors note that the person in front gets almost twice as much attention as the persons on either side. Across all four meetings, we find that the proportions of A^L_i, A^R_i and A^O_i are 17.6%, 16.3% and 21.9% respectively, implying that the likelihood of a subject giving/receiving attention to/from each of the other group members is roughly equal. Therefore, on the basis of the observations made from ground-truth data, we reject ***hypothesis H1, i.e., the person located directly opposite (to the reference subject) does not receive/give significantly more attention***.

Table 3.1: Social attention from ground-truth.

Meeting #	**Subject #**	ST_i **(%)**	A^i **(%)**	$A^i_{(l)}$ **(%)**	A_i **(%)**	AQ_i
1	1	15.2	29.5	24.1	49.3	0.6
	2	44.5	54.6	44.3	57.4	0.95
	3	15.4	42.6	37	71.9	0.59
	4	21.2	36.3	27.6	50.2	0.72
2	1	28.7	57.7	48.2	31.5	1.83
	2	31.1	42	33.6	65.2	0.64
	3	33	42.7	32.9	75.3	0.57
	4	28.2	38.2	29.9	78.9	0.48
3	1	35.4	36.1	25.4	55	0.66
	2	31.2	45.4	39.4	4.3	10.7
	3	15.9	24.3	19.2	45.5	0.53
	4	12.5	25.3	19	58	0.44
4	1	22.7	51	45.9	71.8	0.71
	2	10.4	24.7	20.1	65.8	0.38
	3	16.7	28.1	19.6	56.6	0.5
	4	23	65.3	60.4	57.5	1.14

Validating H2

As can be seen in in Table 3.1 the *attention received*, in general, increases with *speaking time*, and *the speaking activity of a subject influences the amount of attention received by a subject,* ***even when the subject is not speaking***.

Statistically, the correlation between speaking time and attention received is 0.584, which is significant with $p < 0.01$. This corresponds to a coefficient of determination R^2 of 0.341, meaning that speaking time explains 34.1% of the variance in attention received. To conclude, based on the observations made from empirical evidence, ***we validate hypothesis H2, i.e., the overall attention received is influenced by the amount of speech activity***.

Table 3.2: Characterization of meetings based on the variance in speaking time (ST_i) and overall attention received (A^i)

High ST_i, *High* A^i	*High* ST_i, *Low* A^i Meetings 1,3
Low ST_i, *High* A^i Meeting 4	*Low* ST_i, *Low* A^i Meeting 2 (ideal)

Characterizing meetings and persons

Considering the *speaking time* and the overall *attention received* as two dimensions for analysis (Table 3.2), Meeting 2 presents an interesting case where both ST_i and A^i have low variation, showing that all the participants contribute equally while receive roughly equal attention – this corresponds to the *ideal meeting scenario*. Meetings 1 and 3 are cases where the variance along A^i is low and the variance along ST_i is high. Meeting 4 is the opposite: the variance in A^i is high while the speech activity for the various subjects is not very different; *i.e.* someone receives more attention than others, but the speech activity is almost identical across the group. We hypothesize that this corresponds to a group with an established leadership, while the leadership remains undecided in Meetings 1 and 3.

Finally, we define the Attention Quotient for a subject, denoted by AQ_i, as the ratio of the overall attention received to the overall attention given by the individual, *i.e.*, $AQ_i = \frac{A^i}{A_i}$ (Table 1(a)). It has been convincingly shown that meeting behavior can be strongly correlated with one's personality [164]. The 'Mission Survival' data also contains annotated ground-truths for the *Extraversion* and the *Locus of Control* personality traits. *Extroversion* is associated with assertive and highly outgoing personalities while the *Locus of control* (LOC) refers to an individual's nature to be self-determined and undeterred by external factors. In accordance with social psychology literature, we observe a positive correlation between AQ and the *Extraversion* and LOC traits.

Table 3.3: Mean $\overline{ST_i}$, $\overline{A^i}$ and their variance values for the four meetings.

Meeting #	$\overline{ST_i}$ **(%)**	$\overline{A^i}$ **(%)**	var(ST_i)	var(A^i)
1	24.1	40.8	13.90	10.67
2	30.3	45.2	2.23	8.50
3	23.8	32.8	11.25	9.97
4	18.2	42.3	5.95	19.29

The mean ST_i ($\overline{ST_i}$) and A^i ($\overline{A^i}$) values presented in Table 3.3 also agree with the above observations. Meeting 2, which had the lowest variance in *speaking time* and *attention received*, also corresponds to maximum speaking activity and attentional duration, and therefore, corresponds to the ideal meeting scenario. Meetings 1 and 3 correspond to moderate attentional times and speaking activity, while meeting 4 has the minimum speaking activity among all meetings.

3.1.4 Automated Social Attention Estimation

In order to validate the ground truth analysis presented before, we also performed automated analysis of the social attention. The long-term spectral divergence algorithm [231] is used to discriminate between speaking/non speaking regions, while the head-pose-cum-eye center estimation algorithm [285], is employed to estimate the point-of-gaze.The gaze estimation involves integration of a cylindrical head model-based pose estimation and an isophote-based eye-center locator to overcome shortcomings in both systems.

Speaking Activity

In Mission Survival II, speaking activity cues were extracted from close-talk microphone speech signals. The long-term spectral divergence algorithm proposed by [231] was used to discriminate between speaking/non speaking regions. According to this algorithm, in order to detect the vocal activity we assumed that

the most significant information is contained in the time-varying spectral magnitude of the signal.

After segmenting the initial signal, the long-term spectral envelope (LTSE) as well as the long term spectral divergence(LTSD) were estimated, in order to formulate the decision rule for the voice activity detection.

Let *x(n)* be the initial signal, segmented into overlapping frames and *X(k,l)* it's amplitude spectrum for the *k*-th band at frame *l*. The *N*-order LTSE is defined as:

$$LTSE_N(k,l) = max\{X(k,l+j)\}_{j=-N}^{j=+N}$$

The *N*-order long-term spectral divergence between the speech and the noise is defined as the deviation of LTSE with respect to the average noise spectrum magnitude *N(k)* for the *k* band, with *k* = 0,1,...,NFFT-1 , where NFFT is the length of the Fast Fourier Transform. $LTSD_N(l) = 10log_{10}\left(\frac{1}{NFFT}\sum_{k=0}^{NFFT-1}\frac{LTSE^2(k,l)}{N^2(k)}\right)$

The decision rule for the voice activity detection is based on the LTDS between the speech and the noise while the threshold distinguishing speech from non speech regions was adjusted to optimize accuracy as inferred from the ground truth.

The speaking time (ST) for the i^{th} subject is then calculated as:

$$ST_i = \frac{Speaking\ Frames_i}{Total\ Number\ Frames} * 100\%$$

All the meetings have been previously annotated according to the same binary schema, where speaking frames were flagged as "1" while non speaking frames as "0".

The audio cues extracted for all the subjects were fused along with visual cues extracted from head pose and eye gaze, in order to characterize the social behavior of the subjects.

Visual Gaze

To be able to correctly estimate the visual gaze of a user, joint head pose and eye location information should be taken into account. The needed information are retrieved by using the system proposed in [285], which is able to accurately extract head pose and eye center location information from a monocular video sequence.

The method combines a robust cylindrical head model (CHM) pose tracker [312] and an isophote based eye center locator [313], in order to obviate drawbacks of both systems when taken independently. The system integrates the eye locator with the cylindrical head model by using the transformation matrices obtained by both systems in an interleaved way. In this way, the eye center locations are detected given the pose, and the pose is adjusted given the eye center locations.

The 2D eye center locations detected in the first frame are used as reference points. These reference points are projected onto the cylindrical head model and are then used to estimate the following eye center locations. Instead of projecting the found eye center location back to the image plane to obtain a more accurate eye center estimation, the detected eye center locations are projected to a normalized model view. Thus, the displacement between the reference eye point and the current eye center location is independent of the head pose.

The visual gaze of a person indicates the point of interest (i.e. the foveated point) and corresponds to the middle of the visual field of view. Therefore, the displacement of the eyes from their resting position plays an important role in the properties of the visual field of view. Instead of focusing on modeling the shape of the eyes, and how eye movements alter the visual gaze vector in relation to the head position, we assume that extreme eye positions are rare (e.g. the visual gaze vector does not fall outside the visual field of view defined by the head pose only). This assumption allows us to approximate the eye movement as horizontal and vertical shifts on the head surface, which is in turn modeled

as a cylinder. As stated above, the current eye locations and the reference eye locations are used to compute the eye displacement vectors in a normalized (i.e. pose-free) model view. The displacement vectors of the two eyes are averaged and used to modify the basic head-only visual gaze estimate.

Results

Results from automated analysis are presented in Table 3.4, based on which we validate **H3:** ***The use of eye-gaze in conjunction with head-pose improves accuracy of automated social attention estimation***.

Table 3.4: Social attention from automated analysis with head-pose only (HP) and eye-gaze + head-pose ($HP + EG$) information.

Meeting #	**Subject #**	ST_i**(%)**	$A^i(HP)$	A^i **(**HP**+**EG**)**	A_i **(**HP**)**	A_i**(**HP**+**EG**)**
1	1	15.7	33.6	33.4	49.3	49.9
	2	43	46.6	54.4	56.8	57.5
	3	18.2	29.8	33.5	71.1	71.9
	4	20.5	37	40.4	49.9	50.2
2	1	29.3	43.9	46.1	57	55.2
	2	32.5	35.9	39.3	69.7	63.7
	3	33.2	43.9	45.6	90.8	82.3
	4	28	24.4	30.4	63.6	83.9
1	1	32.8	38.6	40.8	47.5	56.8
	2	30.6	39.4	42.8	8.4	7.1
	3	14.7	21.3	21.8	38.4	45.9
	4	13.6	27.6	28	66.8	64.2
1	1	21.3	43.6	45.6	71.9	72.4
	2	13.8	31.9	28	66.1	66.5
	3	18.1	26	25.5	56.7	57.1
	4	22.2	55.8	58	58.4	58.8

We use A^i and A_i estimates to evaluate the accuracy of automated analysis against the ground truth. For A_i, the mean euclidian error between the estimates obtained when using the head pose only (HP) and eye-gaze with head-pose

($HP + EG$) are 7.62 and 5.33 respectively. This corresponds to an error reduction of 30%. For A^i, the corresponding errors are 6.28 and 4.59, yielding an improvement of 26.9%. Therefore, based on the automated social attention estimation results, ***we corroborate H3, i.e, employing eye-gaze along with head-pose cues improves accuracy of automated social attention estimation***.

3.1.5 Conclusions

We have proposed a multimodal framework to analyze social attention in meeting scenarios. To the best of our knowledge, this is one of the first attempts at: (i) simultaneously characterizing both meetings as well as participants by means of multimodal cues, and (ii) explicitly employing eye-gaze as a modality to estimate social attention.

While social attention does not appear to be biased towards the person sitting in front, experimental results indicate a positive correlation between the attention received and the speech activity of a given subject. Importantly, explicitly computing the eye-gaze orientation along with the headpose contributes to a significant reduction in social attention estimation error.

We believe that our results will pave the way for future research connecting social attention with meeting roles and personality traits. To this end, more research is needed to determine the exact relationship between the attention received and given by a person, speech and postural activity, and personality.

3.2 Automatic Modeling of Personality States in Small Group Interactions

It is customary for us to describe people as being more or less talkative, bold or sociable. We employ these descriptors in our daily life to explain and/or predict others' behavior, attaching them to well-known and new acquaintances.

Extraversion, the trait dimension they refer to, is so familiar that we continuously exploit it inconspicuously. Similarly, we talk about other people being more/less prone to anger and frustration (Neuroticism), responsible or attentive (Conscientiousness), and so on. These descriptors relate to traits comprising the well-known Big Five model of personality [54].

The importance of personality for technology and human-computer interaction has also been acknowledged. Studies have shown that personality traits determine people's attitudes toward machines [248] and towards conversational agents [235]. It has been argued that social networking websites could increase the chances of a successful relationship by analyzing text messages and matching personalities [73], and that tutoring systems would be more effective if they adapt to the learner's personality [154].

Moreover, given its relevance in social settings, information on people's personality can be useful for providing personalized support to group dynamics. Several works have explored automated personality analysis [184, 222], often targeting the Big Five model of personality [54] of which Extraversion is a major dimension. The general approach is to isolate promising behavioral correlates of the targeted traits for classification or regression, adopting a thin-slice perspective. In particular, some works have focused on the well-known correlation between Extraversion and prosodic features [222] - higher pitch and higher variation of the fundamental frequency, higher voice quality and intensity - while others such as [184], have also considered verbal cues, including many relating to syntax, content, utterance type, etc. More recently, Lepri lepri2012connectingmedium-grained behaviors enacted in group meetings and related to the social attention (social gaze) in order to automatically predict the Extraversion personality trait.

All these approaches to the automatic recognition of personality have more or less implicitly adopted the so-called *person-perspective* on personality: for a given behavioral sample, classify whether the sample belongs to an extro-

vert or introvert (or equivalently, to a neurotic or an emotionally stable, and so on). The problem with this approach is that it assumes a direct and stable relationship between, e.g., being extravert and acting extravertedly (e.g., speaking loudly, being talkative, etc.). Extraverts, on the contrary, can often be silent and reflexive and not talkative at all, while introverts can at time exhibit extraverted behaviors. Similarly, people prone to neuroticism need not always exhibit anxious behavior, while agreeable people can sometimes be aggressive. While the person-perspective has often dismissed these fluctuations of actual behavior as statistical noise, it has been recently suggested by Fleeson [84] that they are meaningful. The social psychology literature has coined the term personality states to refer to concrete behaviors (including ways of acting, feeling and thinking) that can be described as having a similar content to the corresponding personality traits. In other words, a personality state describes a specific behavioral episode wherein a person behaves more or less introvertly/extravertly, more or less neurotically, *etc*. Personality can then be reconstructed through density distributions over personality states, with parameters such as means, standard deviations, *etc.*, summarizing what is specific to the given individual.

In this work we address the automatic classification of personality states corresponding to the Big Five traits [54], in multi-party meetings. Hence, we will be concerned with classifying whether people behaved: extravertedly/introvertedly; neurotically or in an emotionally stable manner; in a conscientiousness or careless way; agreeably/disagreeably or creatively/uncreatively.

To the best of our knowledge, this is the first work targeting such a qualitative characterization of human behavior in a computational setting, opening new perspectives for the automatic recognition of personality and its relationships to actual behaviors.

3.2.1 Data and Experimental Setup

Our experiments were conducted on 4 meeting videos extracted from the Mission Survival corpus (with a total run-time of 120 minutes). The video stream corresponding to each meeting participant was split into 5-minute long slices, making for a total of 108 clips. Personality state annotation was performed by 30 volunteers (researchers and graduate students) using the Ten Item Personality Inventory [4], a 10-item questionnaire developed to obtain a brief measure of the Big Five dimensions. Annotators were required to assess a participant's personality based on the 5-minute slices containing a close-up view of the subject with the synchronized audio. Contextual information (the behavior of the other meeting participants) was available only through the audio channel. Each video was annotated by three different annotators, and each annotator saw a given subject no more than once. The wordings of the 10 items were modified in order to reflect the different goal of our exercise; hence, rather than asking to assess how much the item "Extraverted, enthusiastic" applied to the subject, it asked about how much it applied to the behavior he/she exhibited in the given audio-video slice. In Table 3.5, we report the global means and standard deviations computed from the annotated personality states. For 16 participants, an average score (5.07 in a range from 1 to 7) for Neuroticism is quite high, while the mean score for Openness (creativity, complexity) is relatively low (3.95).

Table 3.5: Means and standard deviations for personality states.

State	Mean	St. Dev.
Extraversion	4.18	1.66
Agreeableness	4.55	1.05
Conscientiousness	4.99	1.12
Neuroticism	5.07	1.04
Openness	3.96	1.25

In order to perform the classification experiments, scores were quantized

(Low/High) for each personality state by taking the median score as a threshold. Table 3.6 reports the transition probabilities for each personality state. As can be seen, our subjects showed a good degree of consistency in their behaviors: if a person was behaving extrovertedly/introvertedly at time t he/she was more likely to continue behave the same way at t+1. The only exceptions in this regard are (a) the transition from a disagreeable to an agreeable behavior; it appears that participants tend to persist in disagreeable behaviors. (b) Also, a transition from an emotionally stable behavior to a neurotic behavior is as likely as continuing with an emotionally stable behavior.

Table 3.6: Transition probabilities for each personality state.

Extraversion	**L**	**H**
L	.688	.312
H	.296	.704
Agreeableness		
L	.471	.529
H	.259	.741
Conscientiousness		
L	.604	.396
H	.386	.614
Neuroticism		
L	.5	.5
H	.386	.614
Openness		
L	.681	.319
H	.311	.689

3.2.2 Feature Extraction

A total of 37 features, including both low-level features and high-level features, were automatically extracted from the meeting corpus.

Low-level Features

We focused on two classes of features: *Activity* and *Emphasis* [219], measuring vocal signals in social interactions. Activity, implying conversational activity level, usually indicates interest and excitement. It is measured by the z-scored percentage of speaking time (mean and standard deviation of energy per frame, average length in seconds of voiced segments and of speaking segments, fraction of spoken time and voicing rate). For this purpose, the speech stream of each participant is first segmented into voiced and non-voiced segments, and then split into speaking and non-speaking segments.

Emphasis is often considered a signal indicating the strength of the speaker's motivation. Moreover, the consistency of emphasis (lower the variations, higher the consistency) could be a signal of mental focus, while variability may signal an openness to influence from other people. Emphasis is measured by the variation in prosody, i.e. pitch and amplitude. For each voiced segment, the mean energy, frequency of the fundamental formant and the spectral entropy are extracted (mean of formant frequency, confidence in formant frequency, spectral entropy, the value of the largest autocorrelation peak, location of the largest correlation peak, the number of the largest autocorrelation peak, the time derivative of energy in frame). The mean-scaled standard deviation of these extracted values is then estimated by averaging over longer time-periods (standard deviation of formant frequency, confidence in formant frequency, spectral entropy, the value of the largest autocorrelation peaks, the location of the largest autocorrelation peaks, number of the largest autocorrelation peaks, and time derivative of energy in frame).

High-level Features

Social attention features were extracted by jointly processing the audio-video channels using the hierarchical approach developed in [166]: the output of a

Cylindrical Head Model [286] head-pose tracker was used as a proxy for the subject's gaze and finely tuned with the output of a sub-pixel accurate visual-gaze estimation system [287]. Social attention features were defined as the outcome of such joint processing of head-pose and visual-gaze: for each subject, in every frame, Attention Given (to at least one of the other participants) and Attention Received (by at least one of the other participants) were thus available.

The audio channel was processed through use of a Voice Activity Detection system based on the long-term spectral divergence algorithm detailed in [232]. Speaking time was calculated as the percentage of frames in which voice activity was detected over the duration of the processed slice.

For every participant p, the audio-visual features were then fused in order to obtain Attention Given While Speaking (the percentage of p's speaking time during which he/she devotes visual attention to the other members of the group), Attention Given While Not Speaking (the percentage of p's non-speaking time during which he/she devotes visual attention to the other members of the group), Attention Received While Speaking (the percentage of p's speaking time during which he/she receives visual attention from all the other members of the group), and Attention Received While Not Speaking (the percentage of p's non-speaking time during which he/she receives visual attention from all the other members of the group).

3.2.3 Feature Selection

Correlation-based feature selection is a subset selection technique whose objective is to determine the optimal subset of features [108]. This method evaluates the merit of a subset of features computing the individual predictive ability of each feature along with the degree of redundancy between them. The preferred and selected features using this approach are the features highly correlated with the target value and with low inter-correlation values. This method is used in conjunction with a search strategy, typically the Best First search that makes

a search in the space of features subsets, using a greedy hill-climbing with a backtracking facility. The search may start with an empty set of features and search forward (forward search) or with the full set of features and search backward (backward search), or at any point search in both directions, forward and backward. We used the forward search feature selection technique.

The selection has been based on 10-fold cross validation; features selected in at least 8 folds were chosen to train and test the models. The features selected for each personality trait as follows:

1. *Extroversion* – mean of {attention given, attention received, attention received while not speaking, formant frequency, spectral entropy, energy in frame};
2. *Neuroticism* – mean of time derivative of energy in frame;
3. *Agreeableness* – mean of {formant frequency, confidence in formant frequency}, standard deviation of {spectral entropy, number of autocorrelation peaks};
4. *Openness* – mean of {attention received, formant frequency, spectral entropy};
5. *Conscientiousness* – mean of {attention received, attention received while speaking, attention received while not speaking, formant frequency, value of largest autocorrelation peak}.

3.2.4 Automatic Recognition of Personality States

A set of machine learning algorithms, both generative (Nave Bayes, Hidden Markov Models) and discriminative (Support Vector Machines) were used to evaluate the effectiveness of the selected features in modeling the dynamics of personality states. Each algorithm was evaluated in 5 binary classification tasks, one per personality trait.

The leave-one-meeting-out strategy was employed, thus 4 models for each personality state trait were trained on a 3-meetings subset, evaluating them against the remaining one, and finally averaging the results.

The first classifier we applied is Nave Bayes, a simple probabilistic classifier that applies the Bayes theorem and assumes that the presence/absence of a particular feature of a given class (e.g. a personality state) is unrelated to the presence/absence of other features. The main advantage of using Nave Bayes is that it only requires a small amount of training data to estimate the parameters (means and variances) necessary for classification.

As discussed in connection with Table 3.6, there are certain regularities in personality state change that a sequential model might be able to capture. Hidden Markov Models (HMMs) were exploited to this end; HMMs consider the temporal correlation between the samples and define the prior probability of the classes in the current sample as depending upon the posterior probabilities of the classes in the last sample.

More precisely, a left-to-right HMM model for each personality state (one for extrovert/introvert; one for neurotic/emotionally stable; and so on) was represented as follows: t, time; $y(t)$, the feature vector; $x(t)$, the personality state; $p(x)$, the priors for the personality states; $p(x(t)|x(t-1)$, the personality states transitions probabilities; $p(y(t)|x(t))$, the conditional distribution of the observed feature vector given the current personality state at time t. We assumed speaker independence and the feature sequences (one per subject) from all the four meetings were used to train a single HMM. The training was done using the standard Expectation Maximization (EM) algorithm. For prediction, each person is represented by an independent instantiation of the same Markov process. Thus, four independent HMMs represent the four people in a meeting. For classification, we used the standard Viterbi algorithm to compute the most likely sequence of roles.

Finally, we also tested the performance of a powerful discriminative ap-

proach, such as Support Vector Machines (SVMs). The bound-constrained SVM classification algorithms with a linear and with a RBF kernel were used. The cost parameter C and the RBF kernel parameter γ were estimated through the grid technique by means of 10-fold cross validation. Furthermore, the cost parameter C was weighted for each class with a factor inversely proportional to the class size.

3.2.5 Results and Discussion

Table 3.7 below reports accuracy results. Bold figures identify values that are significantly better ($p < .05$) than the relevant criteria, represented by the classifiers that use the priors (binomial tests with Bonferroni correction).

Table 3.7: Personality States Classification Results.

State	NB	HMM	Linear SVM	SVM rbf
Extraversion	**.694**	**.731**	**.648**	**.676**
Neuroticism	**.639**	.574	**.63**	**.62**
Agreeableness	.583	.481	.574	.583
Openness	.5	.5	.546	.583
Conscientiousness	**.583**	.547	.575	.555

Extraversion is the best-recognized behavioral quality: extravert/introvert behaviors are easier (using the considered features) to distinguish than, e.g., Openness. Neuroticism is the second best recognized quality, an interesting result considering that it is obtained by using just one feature, the (mean of the) speech energy derivative. The results for the other states are either non-significant or barely significant (as for Conscientiousness with Nave Bayes).

HMMs yield interesting results only with Extraversion and in this case they perform statistically better than SVM-lin and SVM-rbf. Since, in the ground truth, personality states such as Openness and Conscientiousness show temporal properties similar to those of Extraversion, the ineffectiveness of HMM

to isolate them is probably due to the limited predictive power of the selected features. Finally, SVM classifiers with linear and RBF kernels have similar performance scores.

In general, it seems that the considered non-verbal features have a good power in discriminating between introvert and extrovert behaviors, a result that emphasizes the important role of vocal (e.g. pitch) and social attention features not only for recognizing the Extroversion trait (as widely documented by studies in social psychology and in the automatic analysis of behavior), but also for the recognition of behaviors expected from the trait.

The encouraging results obtained for the Neurotic dimension, employing a single feature is also notable. Putting them together with the results concerning Extraversion, and noting that the two corresponding traits are also referred to as Positive and Negative Affect respectively, it can be concluded that the quality of affect, as it surfaces in actual behaviors, is a most readily detectable characteristic among those constituting the Big Five taxonomy.

We have just begun work concerning a highly complex phenomenon – automatic characterization of the quality of behaviors in terms of personality states. Still, preliminary results appear to have proven its feasibility, opening the way to a new and exciting research area.

Chapter 4

Into the Wild: Implicit Behavioral Patterns Emerging in a Technology-Mediated and Inter-Connected Society

In Section 4.1, we investigate the relationships between social network structure and personality [255]; we assess the performances of different subsets of structural network features, and in particular those concerned with ego-networks, in predicting the Big-5 personality traits. In addition to traditional survey-based data, this work focuses on social networks derived from real-life data gathered through smartphones. Besides showing that the latter are superior to the former for the task at hand, our results provide a fine-grained analysis of the contribution the various feature sets are able to provide to personality classification, along with an assessment of the relative merits of the various networks exploited.

In Section 4.2, we present the SocioMetric Badges Corpus, a new corpus for social interaction studies collected during a 6 weeks contiguous period in a research institution, monitoring the activity of 53 people. The design of the corpus was inspired by the need to provide researchers and practitioners with:

a) raw digital trace data that could be used to directly address the task of investigating, reconstructing and predicting people's actual social behavior in complex organizations; b) information about participants' individual characteristics (e.g., personality traits), along with c) data concerning the general social context (e.g., participants' social networks) and the specific situations they find themselves in.

Using this dataset, in Section 4.3 we turn to tackling research questions as such as *Do we tend to perceive ourselves more creative when surrounded by creative people? Or rather the opposite holds?*

Such information is very valuable to understand how to optimize work processes and boost people's productivity *along with* their happiness and satisfaction. Exploiting real-life data, collected over a period of six weeks in a research institution by means of wearable sensors, in this work we provide insights on human behavior dynamics in the workplace. We explore the use of graphlets, i.e. small induced subgraphs of a network, to encode the local structure of the interaction network of a subject, enriched with affective and personality states of his/her interaction partners. Our analysis shows that graphlets of increasing complexity, encoding non-trivial interaction patterns, are beneficial to affective and personality states recognition performance. We also find that different sensory channels, measuring proximity/co-location or face-to-face interactions, have different predictive power for distinct states.

Finally, in Section 4.4, we investigate a user-centric monetary valuation of mobile PII, in the context of a myriad of mobile apps which collect personal identifiable information (PII) and a prospective market place of personal data. During a 6-week long user study in a living lab deployment with 60 participants, we collected their daily valuations of 4 categories of mobile PII (communication, *e.g.* phonecalls made/received, applications, *e.g.* time spent on different apps, location and media, *e.g.* photos taken) at three levels of complexity (individual data points, aggregated statistics and processed, *i.e.* meaningful interpretations of the data).

In order to obtain honest valuations, we employ a reverse second price auction mechanism. Our findings show that the most sensitive and valued category of personal information is location. We report statistically significant associations between actual mobile usage, personal dispositions, and bidding behavior. Finally, we outline key implications for the design of mobile services and future markets of personal data.

4.1 Inferring Personality Traits from Social Network Structure

The rapid global growth of mobile phone usage has reinforced the need to study the psychological and social implications of this technology. Moreover, recent developments in mobile technologies and the advent of smartphones have sensibly broadened the scope of social sciences' studies: researchers can now exploit data collected by means of such devices, corroborating or even replacing survey-based samplings. Smartphones allow for unobtrusive and cost-effective access to previously inaccessible sources of data related to daily social behavior [230, 157].

Nowadays, these devices are able to sense a wealth of behavioral data: *i)* location, *ii)* other devices in physical proximity through Bluetooth (BT) scanning, *iii)* communication data, including both metadata (logs of who, when, and duration) of phone calls and text messages (SMS) as well as their actual contents, *iv)* scheduled events, *v)* operational status, *vi)* movement patterns, *vii)* usage information, etc.

Recent works have started using smartphone data to automatically infer users' personality traits on the basis of continuously collected data [45, 46, 208]. Chittarajan et al. [45, 46] showed that smartphone usage features (which we will refer to as "actor-based" features from now on, in contrast with "network-based" features) such as the number of calls made or received, their average duration,

the total duration of out/in-going calls, the number of missed calls, the number of unique BT IDs seen, Internet usage, and so on, could be predictive of personality traits. Oliveira et al. [208] investigated also the role played by a limited set of nine structural characteristics of the social networks derived from the rich contextual information available in mobile phone data (call logs). On the other hand, by exploiting survey data, works in the tradition of social psychology and network studies (e.g., Kalish and Robins [143]) have proven the existence of important relationships between individual characteristics and the properties of the networks they are part of and, notably, of the so-called ego-networks.

One important individual characteristic that is expected to influence network size and composition is personality. In Social Psychology it is assumed that people's behavior can be explained to some extent in terms of underlying personality traits, which are seen as enduring dispositions that are relatively stable over time [54]. Talks about personality often refer to several dimensions: we are used to talk about an individual as being (non-)open-minded, (dis-)organized, too much/little focused on him/herself, etc. Several existing theories have formalized these informal ways of approaching personalities by means of multifactorial models, whereby an individual's personality is described through a number of fundamental dimensions known as traits, derived through factorial studies. A well known example of a multifactorial model is the Big Five [135] which owes its name to the five traits it takes as constitutive of people's personality: Agreeableness, Conscientiousness, Extraversion, Neuroticism, and Openness.

Kalish and Robins [143] experimentally examined the effect of individual personality differences on their immediate network environment focusing on ego networks, which consist of a focal node ("ego") and the nodes to whom ego is directly connected to (the so-called "alters") plus the ties, if any, among the alters. Their findings showed that psychological predispositions can explain significant portions of the variance of egocentric network characteristics. In line

with [143], we investigate the hypothesis that individuals' psychological predispositions tend to shape their immediate network environment. In our work, however, we do not exclusively rely on self-reported data, but prominently exploit real behavioral data, collected by means of smartphones, this way taking full advantage of the power of such technology.

Targeting the automatic recognition of Big Five personality traits, our work extends and merges the lines of research followed by Oliveira et al. [208] and Kalish and Robins [143] by: *i)* exploiting both survey and mobile data and comparing the results obtained thereof; *ii)* focusing on several classes of structural network properties (centrality measures, small world and efficiency measures, triadic structures and transitivity measures) and their relationship to personality traits; and *iii)* comparing the results to those obtained from individual activity (actor-based) data.
Our results show that *i)* personality classification from structural network properties compares in a very favorable manner with (and is often superior to) classification by means of individual activity data; *ii)* mobile phones-based behavioral data can be superior to survey ones for the purposes of personality classification; and *iii)* particular feature set/network type combinations promise to perform better with given personality traits.

4.1.1 Related Works

In this section we review key works closely related to ours, from two distinct fields: *i)* social psychology and *ii)* social and ubiquitous computing.

Previous Works in Social Psychology

Traditionally, network theorists devote much of their attention to network structure and how the behavior of individuals depends on their location in the network; for instance, individuals occupying central positions and having denser

or wider reaching networks may gain faster access to information and assistance [29]. Recently, a special interest in the interaction between personality traits and network positions has emerged: personality traits that predispose people to socialize, such as Extraversion or Openness to experience, might foster and accelerate tie formation in social networks while others like Neuroticism might restraint individuals from creating ties. Mehra et al. [190], for example, found that high self-monitors, i.e. people who are concerned about how they are perceived by others, occupied more central positions in the friendship network of a high-tech company.

While these authors applied very specific personality traits, others [143, 150, 238, 270] have addressed more comprehensive instruments such as the five factor model [94]. For example, previous studies demonstrated a positive correlation between Extraversion and ego-network size (e.g., [270]). However, Extraversion tends to decline with age [55] and, after controlling for age, Roberts et al. [238] found no effect of Extraversion on the size of an individual's social network. Klein et al. [150], instead, found that people who were low in Neuroticism tended to have high degree centrality scores in the advice and friendship networks. Unfortunately, their analysis reports only in-degree centrality and hence it does not allow a complete investigation of relationships between the local network structures and the personality traits of the ego. In order to overcome the limitations of this work, Kalish and Robins [143] presented a new method of examining personal networks of strong and weak ties trough a census of nine triads of different types (e.g., WWW, SNS, SSS, where W means "weak tie", S means "strong tie", and N means "no tie"). Their results suggest that people who see themselves vulnerable to external forces tend to inhabit closed networks of weak connections. Conversely, people who seek to maintain their strong tie partners apart tend to be individualists, to believe that they control the events in their lives, and to have higher levels of Neuroticism. Finally, people with strong network closure and "weak" structural holes (where "structural

holes" refers to the absence of ties between parts of the network [36]), tend to be more extraverted and less individualistic.

Previous Works in Social and Ubiquitous Computing

A common characteristic of the works reviewed in the previous section is their being based on information collected by means of surveys (e.g., self-reported social relations). Recently, however, researchers in social and ubiquitous computing have started exploring the wealth of behavioral data made available by smartphones, wearable sensors (e.g., sociometric badges [206]), Facebook [229] and Twitter [93, 228].

Exploiting sociometric badges, Olguin et al. [206] found that Extraversion and Neuroticism were positively correlated with *degree*, *closeness*, *betweenness*, and *eigenvector* centrality measures. Moreover, they found a negative correlation between Conscientiousness and *betweenness* centrality. Gloor et al. [92] found a positive correlation between Openness and Agreeableness on the one hand, and *degree* and *betweenness* centrality on the other. Using Facebook data, Golbeck et al. [93] found a positive correlation between the number of friends (taken as a measure of degree centrality) and Extraversion, and a negative correlation between ego-network *density* and Openness and Extraversion. More recently, Quercia et al. [229] argued that Extraversion is a predictor (albeit weak) for the number of social contacts.

Based on a large dataset consisting of recordings of real-life smartphone usage and personality surveys, Chittaranjan et al. [45, 46] exploited actor-based features (e.g. number and duration of calls, BT hits, etc.) in order to automatically classify personality traits. Their results showed that these features could be predictive of the Big Five personality traits. Moreover, the analysis of these features revealed some interesting trends: extroverts were more likely to receive calls and to spend more time on them, while features pertaining to outgoing calls were found to be not predictive of the Big Five traits. Oliveira

et al. [208] extracted 474 variables from Call Data Records (CDRs), at different time scales, and derived from them the users' social networks; from the latter, they extracted nine structural network features (e.g. *degree*, *efficiency*, etc.). For three personality traits (Extraversion, Agreeableness, and Openness), they obtained significant improvements in classification performance when using some of these structural network characteristics. Inspired by Oliveira et al. [208], our work extends the number and types of global and local social network structural properties to include centrality, small world and efficiency measures, triadic structures and transitivity measures.

4.1.2 Dataset

For our work we exploited a dataset capturing eight complete weeks in the lives of 53 subjects living in a married graduate student residency of a major US university, collected between March and May 2010. Each participant was equipped with an Android-based cell phone incorporating a sensing software explicitly designed for collecting mobile data. Such software runs in a passive manner, and does not interfere with the normal usage of the phone [6].

The data collected consisted of: *i)* call logs, from which we built a Call network whereby participants act as nodes and the numbers of calls between two nodes as edge weights, according to the method used in Eagle et al. [75]; *ii)* proximity data, obtained by scanning near-by phones and other Bluetooth (BT) devices every five minutes, which allowed us to build a BT proximity network with, again, participants as nodes and the counts of social interactions derived from BT data as edge weights; *iii)* data from a survey administered to participants, which provided self-reported information about personality (Big Five) and relationships among subjects. Concerning the latter, the participants were required to assess their closeness to each other on a "0 (no close at all) to 10 (very close)" scale. This information was used to build the Survey network using the obtained scores as edge weights.

More specifically, social interactions were derived from Bluetooth proximity detection data in a manner similar to those in previous reality mining studies [76, 179]. The *Funf* phone sensing platform[1] was used to detect Bluetooth devices in the user's proximity. The Bluetooth scan was performed periodically, every five minutes, in order to keep from draining the battery while achieving a high enough resolution for social interaction. With this approach, the BT log of a given smartphone would contain the list of devices in its proximity, sampled every 5 minutes. Knowing the BT identifiers of each smartphone in the study, we could thus infer when 2 participants' phones were in proximity.

The number of subjects varies from one network to another, due to several factors. For instance, some subjects were isolates in the Call network: this could derive from the fact that they had only called people not participating in the data collection (but our data did not include such external calls); or, the call logging might have suffered malfunctions. Thus, these subjects were discarded from the Call network, under the assumption that their empty ego-network structure would introduce undesired noise for the purpose of personality classification. Beside the two basic behavioral networks (Call and BT) and the one based on survey data, we formed a complex behavioral network by combining Call and BT networks in such a way that its node set was the intersection of BT and Call networks' node sets and its edge weights were a linear combination (the sum of the normalized edge weights) of BT and Call networks' weights. All our four networks are undirected; they are quantitatively summarized in Table 4.1.

	Number of nodes	**Number of edges**
Call net	44	77
BT net	50	823
Call∩BT net	42	609
Survey net	53	590

Table 4.1: Quantitative summary of the four networks under analysis.

[1] http://funf.org

Big Five personality traits were measured by asking subjects to use 1-5 point scales to answer the online version of the 44 questions Big Five questionnaire developed by John et al. [135].

Trait	Description
Agreeableness	sociable, assertive, playful
Conscientiousness	self-disciplined, organized
Extraversion	friendly, cooperative
Neuroticism	calm, unemotional
Openness	creative, intellectual, insightful

Table 4.2: Big 5 personality traits explained.

The Big Five questionnaire owes its name to the five traits, explained in Table 4.2, that it takes as constitutive of people's personality.
The scores of the five traits were computed by summing the (inverted when needed) raw scores of the items (i.e. questions) pertaining to each trait. The results (average, standard deviation, median, minimum and maximum values) are reproduced in Table 4.3. We performed a Lilliefors' goodness-of-fit test of com-

	Mean	St.Dev.	Median	Min	Max
Agre.	34.25	5.03	34	21	45
Cons.	32.49	5.5	34	20	42
Extr.	26.15	6.78	25	12	39
Neur.	22.32	5.85	23	9	34
Open.	33	6.87	34	11	45

Table 4.3: Statistics for the Big 5 personality traits.

posite normality on each trait's distribution. All traits are normally distributed ($p < 0.05$).

4.1.3 Extraction of Network Characteristics

Drawing on previous works, we derived a set of network characteristics describing our networks.

Centrality measures	Degree/Closeness/Betweeness Centrality [89] Eigenvector/Information Centrality [28, 162]
Efficiency measures	Nodal/Local Efficiency [161] Mean Nodal/Local Efficiency*
Transitivity and triadic measures	Global/Local Transitivity* [299] Mean Local Transitivity* Triads {1, 3, 11, 16}* [63] Triads {WWW, SSS, WNW, WSW, SNS, SNW, SWS, SWW, SSW}* [143]

Table 4.4: Extracted network features (* indicates computation performed on the ego-net).

The features reported in Table 4.4 have been extracted from both weighted and unweighted networks when applicable.
In the following subsections we describe in detail and justify the features extracted. Previous findings on the relationship between personality traits and structural network properties are reported in Table 4.5.

	Degree		Closeness		Betweenness		Eigenvector	Transitivity	WWW triads		SSS triads		SWS triads
	+	-	+	-	+	-	+	+	+	-	+	-	-
Agre.	[63, 92, 150]				[63, 92]			[189]					
Cons.	[144]	[302]		[302]		[206, 302]		[189]					
Extr.	[93, 144, 302]		[302]		[302]		[302]			[143]	[143]		
Neur.	[302]	[144, 150]	[302]		[302]		[302]		[143]			[143]	[143]
Open.	[92]	[144, 150]			[92]								

Table 4.5: Previously found relations between network measures and personality (+/- indicate positive/negative correlations).

Centrality Measures

In the literature there is ample, though not always converging, evidence of a relationship between centrality measures and Big Five traits. For instance, according to Kanfer and Tanaka [144] all the Big Five personality traits, with the exception of Agreeableness, correlate closely with *degree*, and more precisely with *in-degree*; moreover, agreeable persons tend to occupy central positions and report more interacting with others while outgoing (extraverted) and secure (low Neuroticism) subjects had more people reporting interacting with them. Klein et al. [150] found negative correlation between *in-degree* centrality from Neuroticism and Openness, and a positive effect of Agreeableness in friendship networks of work group members. Surprisingly, Extraversion had no effect on friendship centrality.

According to [302], Conscientiousness negatively correlates with *closeness*, *betweenness* and *degree* centrality; Extraversion and Neuroticism (the latter in a less evident manner) positively correlate with *degree*, *closeness*, *betweenness* and *eigenvector* centrality. Olguin et al. [206] obtained evidence for the negative correlation of Conscientiousness and *betweenness* centrality. In a more recent study conducted by Gloor et al. [92], the authors found significant positive correlations between Openness and Agreeableness and *degree* and *betweenness* centrality.

Inspired by these previous works, we extracted the three standard measures of centrality proposed by Freeman: *degree*, *betweenness*, and *closeness* centrality [89].
These centrality measures can be divided in two classes: those based on the idea that the centrality of a node in a network is related to how close the node is to the other nodes (e.g. *degree* and *closeness* centrality), and those based on the idea that central nodes stand between others playing the role of intermediary (e.g. *betweenness* centrality).

A different interpretation of centrality is given, for instance, by delta centrality measures, which take into account the contribution of a node to network cohesiveness, inferred from the observed network variation when the node is deleted. We computed a delta centrality measure recently proposed by Latora and Marchiori [162]: *information* centrality, based on the concept of efficient propagation of information over the network [160, 161].

Another centrality measure we extracted, *eigenvector* centrality [28], accords to each node a centrality score depending both on the number and the quality of its connections: having a large number of connections is still valuable, but a vertex with a smaller number of high-quality contacts may outrank one with a larger number of mediocre contacts.

Small World and Efficiency Measures

Latora and Marchiori's concept [160, 161] of *efficiency* can be used to characterize how close to a *small world* a given ego-network is.

Small world networks are a particular kind of networks that are highly clustered, like regular lattices, and have short characteristic path lengths, like random graphs [301]. The *efficiency* E of a graph G containing N nodes is defined as:

$$E(G) = \frac{1}{N(N-1)} \sum_{i \neq j \in G} \frac{1}{d_{ij}} \tag{4.1}$$

where d_{ij} is the shortest path length between two nodes i and j in graph G. The use of efficiency measures for the performance evaluation of structural network features is justified by the hypothesis [178] that the rate at which information flows within an ego-network is influenced to some degree by the personality of the ego.

For each node $i \in G$, *local* efficiency is defined as [161]:

$$E_{\text{loc}} = \frac{1}{N} \sum_{i \in G} \frac{E(G_i)}{E(G_i^{\text{ideal}})} \tag{4.2}$$

Here, the normalization factor $E(G_i^{\text{ideal}})$ represents the efficiency of the ideal case G_i^{ideal} in which i's ego-network G_i has all the $k_i(k_i-1)/2$ possible edges, where k_i is the number of edges incident with i. E_{loc} is an average of the *local* efficiency and plays a role similar to transitivity, which will be treated in the next section. Since $i \in G_i$, the *local* efficiency E_{loc} tells how efficient the communication is between i's neighbours when i is removed; in other words, *local* efficiency gives a measure of the response, in terms of efficiency, of i's ego-network when i is removed.

Conversely, *nodal* efficiency is defined as the inverse of the harmonic mean of path length, hence for a given node $i \in G$ it is calculated as:

$$E_i^{\text{nodal}} = \frac{1}{N-1} \sum_{j \in G} \frac{1}{L_{ij}} \tag{4.3}$$

We computed *nodal* and *local* efficiency for each node in the networks, along with the *mean nodal* and *mean local* efficiency of each ego-network. All were extracted both on the weighted and unweighted networks, accounting for a total on eight efficiency measures computed.

Transitivity Measures

In [302], extraversion was found to negatively correlate with *local transitivity*; McCarty and Green [189] found that agreeable and conscientious persons tend to have well-connected networks. To account for a possible contribution of this notion to personality prediction, we computed the following three transitivity features: *i) global* transitivity of the ego-networks, *ii) local* transitivity, and *iii) mean local* transitivity.

Global transitivity gives an indication of clustering properties at the level of the entire ego-network. It is based on triples' counts, where a triple is a set of three nodes connected by either two (open triple) or three (closed triple) ties. The *global* transitivity of a given graph G is then defined as the ratio between

the number of closed triples in G and the total number of triples. For each ego-network, we computed this measure, which gives an indication of the clustering in a network, and is often referred to as clustering coefficient.

The *local* transitivity of a node, in turn, measures how close its neighbors are to forming a clique (i.e. a complete graph) and the graph to a small-world network [301]. For a node i, *local* transitivity is defined as the proportion of ties between the nodes in i's neighborhood (i's ego-net) to the number of ties that could possibly exist between them.

Finally, we computed i's *mean local* transitivity as the mean of the local transitivity values of i's adjacent nodes.

Triadic Measures

In [63] each triad is described by a string of four elements: the number of mutual (complete) dyads within the triad; the number of asymmetric dyads within the triad; the number of null (empty) dyads within the triad; and, finally, a configuration code for the triads which are not uniquely distinguishable by the first three elements. In the case of directed graphs, every triad may thus occupy one of the 16 possible distinct states. Conversely, in the case of undirected networks, as in our experiments, the triadic census reduces to the following four states: *i)* Triad_1, the empty triad; *ii)* Triad_3, the ratio of triads where two nodes are connected; *iii)* Triad_11, the ratio of triads where a given vertex i is connected to the node j and the node z and there is no edge between the latter two; *iv)* Triad_16, the ratio of triads representing the complete graph, namely i is connected with j and z, and j and z are also connected.

Recently, Kalish and Robins [143] argued that triad proportions can provide more accurate and informative depictions of the egocentric networks than global measures. They also argued that those ego-network properties are significantly associated with the ego's personality traits. In details, when ego is connected to two alters, $alter1$ and $alter2$, the triad that depicts the relation-

ship between these three actors is denoted by a three letter combination. The first letter indicates the strength of tie between ego and $alter1$ (S or W, for Strong or Weak tie), the second letter the strength of the tie between alters (S, W, or N, for Strong, Weak, or No tie) and the third letter the strength of the tie between $alter2$ and ego (S or W). Given the symmetry of triads, $alter1$ and $alter2$ are interchangeable and SNW and WNS are thus same triad; hence, a total of nine possible triads can occur in egocentric networks: SSS, SWS, SNS, WWW, WSW, WNW, SSW, SWW, SNW. As for [63], the census is not just a count of the different triad types but rather the proportions of each type against the total number of possible triads given the number of alters in the network: in this way, egocentric networks of different sizes can be compared.

Among the nine triads defined by Kalish and Robins, SSS and WWW represent strong and weak tie network closure, respectively, while WNW, SNS, and SNW represent three different types of structural holes. In particular, WNW represents weak structural holes as implied by Granovetter [99]; SNS represents strong structural holes as permitted by Burt [36]; and SNW represents a mixed structural hole between a strong and weak tie. This third structural hole is permitted by Burt [36] but is also implied by Granovetter [99].

Exploiting this typology of triads, Kalish and Robins suggested that Neuroticism is positively associated with the proportion of WWW triads and negatively associated with the proportion of SSS and SWS triads. Conversely, they found Extraversion to be negatively associated with the proportion of WWW triads and positively associated with the proportion of SSS triads.

In our case, the definition of Strong and Weak ties was established as follows: following [143], from the weighted adjacency matrix, we used as a threshold the 59th percentile of the edge weights array cumulative distribution; then, edges with a weight higher or equal than that threshold were considered as S (Strong) while edges weighting less than the threshold were marked as W (Weak).

4.1.4 Automatic Prediction of Personality Traits

We turn now to investigating the predictive power of the different features sets discussed above by comparing the results obtained on a personality classification task. To this end, personality traits scores were quantized into two classes (Low/High), using the median values reported in Table 4.3. Classification was performed by means of Random Forests ensemble classifiers [33]. We chose Random Forests because they satisfy the max-margin property, they do not require parameter tuning, and, importantly, they are feature-space agnostic, *i.e.* they do not require the specification of a feature-space, as in Support Vector Machines (SVMs) do through the kernels. Moreover, Random Forests are one of the most accurate learning algorithms available [40]. We ran the same experiments described below also by using SVMs with linear and RBF kernels and obtained less stable and accurate results.

The five sets of features introduced above were exploited and compared: *i)* centrality measures, *ii)* efficiency measures, *iii)* Davis & Leinhardt's triad census [63], *iv)* Kalish & Robins' triad census [143], and *v)* transitivity measures. To them, we added three more sets of features, consisting of: *vi* centrality and efficiency features together – i.e. the union of *i)* and *ii)*; *vii)* all the triadic measures – i.e the union of *iii)* and *iv)*; and *viii)* all the features assessing local connectivity – i.e. the union of *v)* and *vi)*. The resulting 8 sets of features were computed on the Survey, Call, BT and on the compound Call∩BT networks described above.

Classification results were validated by embedding bootstrap [153] in a Leave-One-Out strategy as follows: first, a new dataset D was generated by leaving subject i out of the original dataset; then, for 100 iterations, a new training set was created by randomly sampling D (with replacement) and used to train a classifier, the latter being tested on the left out instance i. As a baseline, we chose the classifier that always outcomes the majority class (e.g. in case of

perfect balance, the baseline's accuracy is 50%). The obtained mean accuracy values are reported in Tables 4.6-4.10; each table addresses one of the Big Five trait, with columns distinguishing the results according to network type and marginals indicated in italics. As can be seen, in all cases the performances are well above those of the baseline.

	Survey	**BT**	**Call**	**Call∩BT**	
baseline	50.94	58	56.8	52.3	
Centrality measures	65.12	73.59	68.82	62.21	*67.57*
Efficiency measures	67.14	67.14	72.99	61.88	*67.34*
Centrality + Efficiency measures	64.09	71.59	69.87	63.75	*67.34*
Transitivity measures	59.86	73	58.18	66.23	*64.36*
Kalish & Robins' triads [143]	65.56	70.2	70.34	72.16	*69.37*
Davis & Leinhardt's triads [63]	60.24	70.24	61.15	67.49	*64.71*
All triads	64.45	71.62	69.97	69.28	*68.71*
All triads + Transitivity measures	64.56	71.19	68.98	67.9	*68.08*
	63.88	*71.07*	*67.54*	*66.36*	

Table 4.6: Accuracies on Agreeableness, and *marginals*.

Accuracy figures were converted into global ranks and an all-encompassing analysis of variance on ranks was ran with design Trait(5)*Network-Type(4)*Feature-Set(8). All the various main and interaction effects turned out to be significant ($p<.05$). Pairwise comparisons (with Bonferroni adjustment for multiple comparison and overall α=0.05) on marginal means for Network Type reveals the following ordering of accuracy values: BT (68.56) $>$ Call∩BT (66.88) $>$ Survey (65.54) $>$ Call (62.86). Concerning the feature sets, the same procedure revealed that both Centrality (67.40) and Centrality+Efficiency (66.97) outperform Davis & Leinhardt's triads (64.74) and Transitivity (65.20). Finally, the marginal accuracy values for the Big Five traits could be ordered as follows: Openness (68.23) $=$ Agreeableness (67.21)$>$ Extraversion (65.77) $=$ Conscien-

	Survey	**BT**	**Call**	**Call∩BT**	
baseline	54.7	52	56.8	59.5	
Centrality measures	67.07	72.25	63.83	76.97	*69.88*
Efficiency measures	65.26	67.02	58.82	66.62	*64.53*
Centrality + Efficiency measures	67.41	69.87	63.41	74.43	*67.93*
Transitivity measures	67.02	62.08	62.42	65.66	*64.34*
Kalish & Robins' triads [143]	65.84	65.85	58.88	64.12	*63.37*
Davis & Leinhardt's triads [63]	67.12	66.05	56.84	69.14	*64.89*
All triads	65.94	67.12	57.35	64.81	*64*
All triads + Transitivity measures	65.52	67.03	56.59	64.98	*63.72*
	66.06	*67.16*	*59.52*	*68.34*	

Table 4.7: Accuracies on Conscientiousness, and *marginals*.

	Survey	**BT**	**Call**	**Call∩BT**	
baseline	54.7	60	54.5	54.8	
Centrality measures	66.67	73.08	59.45	68.2	*67.02*
Efficiency measures	62.53	70.96	58.92	69.62	*65.5*
Centrality + Efficiency measures	66.27	71.86	61.82	67.7	*67.03*
Transitivity measures	61.76	79.74	51.55	66.95	*65.3*
Kalish & Robins' triads [143]	64.7	70.3	59.85	69.01	*66.01*
Davis & Leinhardt's triads [63]	58.45	77.61	51.63	64.87	*63.36*
All triads	63.47	73.78	59.42	68.53	*66.38*
All triads + Transitivity measures	64.56	74.89	58.7	67.98	*66.69*
	63.55	*74.03*	*57.67*	*67.86*	

Table 4.8: Accuracies on Extraversion, and *marginals*.

tiousness (65.27) > Neuroticism (63.24). In summary, at the global level BT is the most, and Call the least, efficient Network for classification purposes; Centrality features (be they alone or in conjunction with Efficiency one) outperform

	Survey	**BT**	**Call**	**Call∩BT**	
baseline	52.8	60	59.1	54.8	
Centrality measures	62.87	60.54	73.74	63.82	*64.99*
Efficiency measures	64.99	60.6	66.82	63.21	*63.86*
Centrality + Efficiency measures	62.91	58.53	72.99	63.39	*64.21*
Transitivity measures	59.25	64.8	57.03	64.02	*61.26*
Kalish & Robins' triads [143]	66.67	61.63	59.62	69.75	*64.38*
Davis & Leinhardt's triads [63]	59.37	60.19	62.3	61.59	*60.76*
All triads	64.84	59.59	60.91	66.58	*62.92*
All triads + Transitivity measures	64.17	59.96	61.51	65.35	*62.7*
	63.13	*60.73*	*64.37*	*64.71*	

Table 4.9: Accuracies on Neuroticism, and *marginals*.

	Survey	**BT**	**Call**	**Call∩BT**	
baseline	50.9	54	56.8	52.4	
Centrality measures	70.71	65.56	68.39	65.13	*67.57*
Efficiency measures	71.2	69.44	66.84	66.31	*68.63*
Centrality + Efficiency measures	70	69.79	68.31	64.02	*68.22*
Transitivity measures	73.52	70.44	63.37	77.05	*71.13*
Kalish & Robins' triads [143]	69.66	68.47	63.21	61.26	*65.98*
Davis & Leinhardt's triads [63]	70.82	70.32	63.82	75.52	*70.1*
All triads	70.69	70.32	64.03	62.85	*67.3*
All triads + Transitivity measures	72.01	71.71	63.53	64.97	*68.39*
	71.07	*69.51*	*65.19*	*67.14*	

Table 4.10: Accuracies on Openness, and *marginals*.

Davis & Leinhardt's triads and Transitivity while Agreeableness and Openness are the traits that are best recognized. This global picture misses many interesting details, to which we now turn by discussing the results of analyses of

variance, one per trait, that were run with Feature-Set and Network-Type as factors, in a 8*4 design. The results are summarized in Table 4.11. Accord-

	Agre.	**Cons.**	**Extr.**	**Neur.**	**Open.**
Network-Type	11.422***	17.113***	44.254***	4.082**	7.199***
Feature-Set		3.633**			2.124*
Network-Type * Feature-Set	1.699*		1.412!	2.269**	1.529*

Table 4.11: ANOVA results. F values and their significance values – !: $\mathbf{p < .1}$; *: $\mathbf{p < .05}$; **: $\mathbf{p < .01}$; ***: $\mathbf{p < .001}$.

ing to Table 4.11, the network type has a significant influence on classification results for all traits. A detailed analysis of the sources of these effects (same procedure as above for pairwise comparisons) yields the following patterns:

- *Agreeableness* – BT > Call = Call∩BT > Survey (see Table 4.6);
- *Conscientiousness* – BT = Call∩BT = Survey > Call (see Table 4.7);
- *Extraversion* – BT > Call∩BT > Survey > Call (see Table 4.8);
- *Neuroticism* – no clear ordering, though BT is significantly worse than Call and Call∩BT;
- *Openness* – no clear ordering, though Call is worse than BT and Survey.

In summary, the neat ordering among network type that we detected at the global level is substantially confirmed at the level of the single trait: in all but one case (Neuroticism), BT is the best performing network and Call is the worst one.

Turning to the Feature-Set effects of Table 4.11, they are significant only with Conscientiousness and Openness. With the former, the effect is due to the better accuracies of Centrality features with respect the other feature sets; with Openness, the joint analysis of the Feature Set main effect and of the Network Type*Feature Set interaction reveals that the (otherwise quite low) performances of Call∩BT significantly increase when Davis & Leinhardt's triads or

Transitivity are used. The remaining two interaction effects concern Agreeableness and Neuroticism: the former can, at least in part, be attributed to a performance drop of Call with Davis & Leinhardt's triads and Transitivity, see Table 4.6; the second interaction effect can be traced back to the accuracy increase obtained when Centrality and Centrality+Efficiency are computed from Call. We also discuss the interaction effect for Extraversion: though only marginally significant ($p<.1$), it is worth commenting because it highlights opposing patterns between BT and Call networks, with the former sensibly increasing its performance with Davis & Leinhardt's triads and Transitivity and the latter decreasing when the same feature sets are exploited.

We see, therefore, that the pattern highlighted above when discussing the effects of Feature Set at the global level, stems from specific interactions among Network Type, Trait, and Feature Set. Moving from coarser to finer grained considerations, the survey network never outperforms the other network types (though it provides very good results with Openness), suggesting that, despite the many problems that might affect them (e.g., sparseness and incompleteness in the case of the Call network), behavioral data are in a better position than survey data for automatic personality prediction purposes. The second point concerns the relationships between network types and feature sets: in general, Call's performances tend to decrease with the various types of transitivity and triadic features; BT performance, in turn, are more stable (and higher) across features sets and personality traits. The results from the trait-specific ANOVAs allow refining these general associations: Centrality computed on Call yields high performances with Neuroticism; Davis & Leinhardt's triads and Transitivity computed on BT improve the classification accuracy with Extraversion, and they do the same for Openness when computed on Call$\cap$BT. The improved results for Extraversion with Davis & Leinhardt's triads can be further analyzed in the lights of the correlations between the relevant features and the Extraversion trait in the BT network (henceforth, we discuss only correlation with signif-

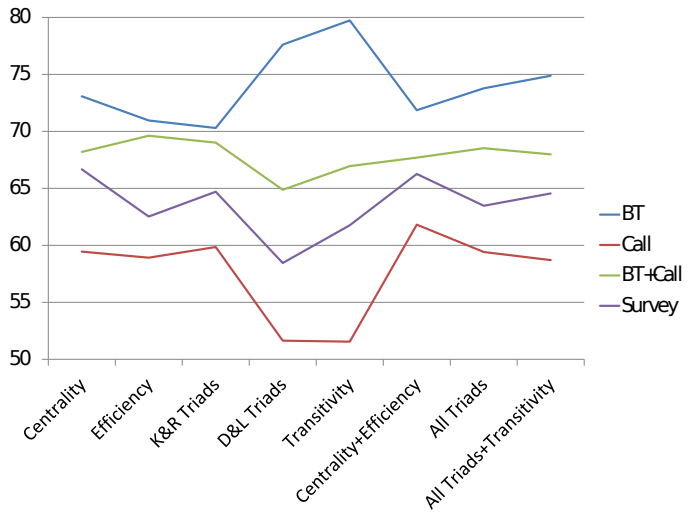


Figure 4.1: Feature sets' performance on Extraversion.

icance level p<.01). In particular, we found a positive association (.281) for Triad_16, the (ratio of) triads representing a complete graph, and negative associations (-.345, -.283 and -.237, respectively) for Triad_1, the empty triad, Triad_3, the triads with two connected nodes, and Triad_11, the triads with two edges. In other words, in our BT network, extraverts tend to have more complete triads and less incomplete or empty ones, than introverts. As one may put it, extraverts seems likely to keep their close partners together, perhaps by actively seeking to introduce them to one another at the social gatherings captured by the BT network. Kalish and Robins' triads, which consider the strength of the ties, seem to be slightly less informative for Extraversion classification, according to our results; and, correspondingly, we find only a couple of significant correlations in this case: with WNW (-.228) and with SNW (-.235) triads. In these respects, our BT network does not replicate Kalish and Robins [143]' findings that extraverts have proportionally more SSS and SWS triads and fewer WWW triads, a difference that can reasonably be due to the different types of networks these data are drawn from (survey data in [143], BT in the present discussion). Finally, all features in our transitivity set significantly correlate with

Extraversion: local transitivity (.301), mean local transitivity (.282) and global ego-network transitivity (.285); in correspondence, classification performance on the BT network gets up to 80%. It should be noticed that our correlation figures contrast with those obtained by Wherli [302] where Extraversion was found to negatively correlate with local transitivity. However, our results seem supported by those of Hallinan and Kubitschek [109] who, examining the relationship between tolerance for intransitivity and friendliness, found that friendly students have a lower tolerance for intransitive triads and tend to remove them over time. Finally, no significant correlations could be found in the Call network between the Extraversion trait and any of the features composing the Davis & Leinhardt's triads and the Transitivity sets, possibly explaining the drop in accuracy discussed above.

Turning to Neuroticism, the association with centrality measures in the call network that our classification results reveal can be traced back to the level of correlations to degree centrality (.257), a datum that is in line with findings in [302]. By indicating a more substantial (though not necessarily linear) relationship between centrality features and Neuroticism, our classification results call for further investigation of the underlying phenomena.

Our conclusions concerning Agreeableness are similar to those for Neuroticism. In the literature, this trait has not been investigated much by means of network-level measures. On our side, we could only find a significant positive association in the Call network between Agreeableness and local efficiency (.246), which measures the mean efficiency internal to an ego-network, an index related to small world formation. Correspondingly, the Call network accuracy gets up to 73% with the Efficiency feature set. As it seems, therefore, more agreeable people have some tendency towards forming small worlds than less agreeable ones; again, this is a datum that, together with its import to the explanation of our classification results, needs further investigation. The literature does not offer much to discuss, and compare with, concerning the elusive trait

of Openness. Given this lack of information, the level and type of recognition accuracy that we obtained is even more remarkable: at the global level, Openness is one of two best recognized trait, with average accuracy 68.23; it seems capable of taking specific advantage of features, such as Davis & Leinhardt's triads and Transitivity, which measure the level of connectedness of the egonet; it is also the trait were information from the surveys performs better. Definitely, more work is needed.

4.1.5 Discussion and Comparison with Previous Works

A number of recent works have used mobile phones data in order to automatically infer and predict personality traits ([45, 46, 208]). In particular, [45] exploited actor-based features (e.g., the number of calls made or received, their average duration, the total duration of out/in-going calls, the number of missed calls, the number of contacts associated with missed called, the number of unique BT IDs seen, Internet usage, and so on). In this work, we have focused on the usage of network-level features, arguing that they are more informative for the task at hand than actor-based ones. In order to contribute to the assessment of the relative merits of these two approaches, we computed actor-based features from our networks and compared the results obtained through them to those discussed above. Because of the different nature of our data, we could not fully replicate Chittaranjan et al.'s study as we only had available the following activity level features: number of outgoing calls, number of incoming calls, number of calls from a unique subject, number of calls directed to a given subject, number of unique subjects in proximity (through BT), max time a subject was seen in proximity, total time seen in proximity. The obtained average accuracies are reported in Table 4.12.

While consistently performing above the baseline, actor-based features seem to perform worse than network-based ones with any traits, with the possible exception of Conscientiousness, as a comparison between Table 4.12 and Tables

4.6-4.10 above shows. An approach more similar to ours is reported in [208].

	Agre.	**Cons.**	**Extr.**	**Neur.**	**Open.**
Call features only	71.2	65.4	72.9	66.1	69.2
BT features only	67.1	67.7	62.9	63.2	67.1
All (Call + BT) features	69.1	65.7	68.8	65.2	69.8
baseline	50.9	54.7	54.7	52.8	50.9

Table 4.12: Accuracies obtained using actor-based features.

In this work, the authors used 9 network-based features including the number and the weight (measured by the number of reciprocal phone calls) of contacts (degree of the nodes), the number and social distance between relevant contacts, etc. The authors reported significant improvements ($p<.05$) in the classification performance for 3 traits, Extraversion, Agreeableness, and Openness, when the network-level features were included among predictors. Although at a first sight in line with the general trends we obtained concerning those three traits, more direct comparisons are made difficult by the limited amount of information about the way the network-level features were computed and used in [208].

More generally, with respect to the quoted studies, the present work has addressed a larger number of aspects relevant to the usage of behavioral data from mobile phones for the task of automatically predicting personality traits. In the first place, we emphasized the importance of egonets' structural characteristics, as those that more clearly associate with personality traits variations. In the second place, we systematically investigated the predictive power of those structural properties with networks arising from different behaviors (call logs and BT proximity), and compared the obtained results to those with survey data, which still are the most common source of data in the social sciences. The results we obtained provide encouraging evidence that behavioral data are indeed better suited to our task. Moreover, the detailed analyses conducted on the relation-

ships among the feature set exploited, the different networks and the Big Five traits have allowed us to reveal their relative merits for the task of personality prediction. Among the relevant findings, we list the superiority of triadic and transitivity features for Extraversion prediction on BT data; the importance of centrality and efficiency features from Call data for Neuroticism and Agreeableness; the overall greater richness of the information provided by BT data with respect to Call data.

4.1.6 Practical Implications and Limitations

The automatic determination of personality from mobile phone data can provide a new and interesting framework for mobile and, more in general, pervasive computing. As observed in [46], the ability of inferring and predicting personality and other psychological variables through contextual data collected by mobile phones could be used in various ways in the context of mobile applications.

In the first place, previous works have shown that personality is linked to user interface preferences [34]. The personality of a user might also determine the kind of functions he/she is disposed to use on the phone, as in the case of recommendation systems that attempt to match the preferences and personalities of their users [98].

Another important practical implication of our research program is the use of the automatic understanding of personality from mobile phone data for the design of more effective strategies of mobile persuasion. Given their pervasiveness, mobile phones are becoming the most powerful channel for persuasion, more influential than TV, radio, print, or the Internet [87]. At the same time, some studies (e.g., [106]) have convincingly shown that psychological variables affect if and how people are amenable to persuasion as well as the choice of the best means to bring it about; as a consequence, automatically inferred personality traits can be used to build more effective change-inducing systems.

Turning to the limitations of the present study, we list the following ones: the relatively small size of the sample; the fact that it comes from a population living in the same environment (our subjects were all married graduate students living in a campus facility of a major US university); the non-availability of behavioral data concerning the interaction with people not participating in the data collection, a fact that is common to many other studies of this type and that has been also pointed out by [229]. The first two problems are at least partially attenuated by the large variability of the sample in terms of provenance and cultural background, which can be expected to correspond to a wide palette of interaction behaviors that efficaciously counterbalance the effects of sample's small size and of living-place homogeneity.

4.1.7 Conclusions

This paper aimed to contribute to advance the state of the art in the automatic analysis of people personality by deepening and extending previous works along several dimensions: a) the usage of network-level features (and in particular those addressing the properties of *egonets*) and the comparison of the results obtained with those attained through actor-based features; b) the comparison of the results from two different types of mobile phone data with those based on more traditional surveys; c) the systematic analysis of the relative strengths and weaknesses of the exploited feature sets across network types and the Big Five traits. Despite the limitations of this study discussed above, we believe that our results have provided compelling evidence that mobile phones-based behavioral data can be superior to survey ones for the purposes of personality classification and that egonet-based features can improve performance over actor-based ones. Moreover, we have provided many new insights concerning the feature set/network type combinations that promise to perform better with given personality traits.

4.2 A Multilevel Behavioral Dataset for Social Behavior in Complex Organizations

Studying organizational behavior over extensive periods of time has long been a challenge in social science [12]. Human observers have been employed in the past, but their observations are subjective and it is difficult for them to remain unobtrusive in an organizational environment. In addition, employing a large number of these observers for more than a short period of time is prohibitive in terms of costs. Surveys have been used extensively, but these too suffer from subjectivity and memory effects [149].

To mitigate some of these problems, e-mail, blogs, wikis, and more generally electronic communication have recently been employed to examine relationship structures (i.e. social network structure) [12]. However, while digital communication is important in the modern workplace, face-to-face interaction still represents a large and important share of organizational communication, information exchange, socialization and informal coordination [307].

Until recently, the prevalence of data from electronic communication (or any other systems exploited by people) was justified by the difficulties encountered in the collection of data concerning face-to-face communication at the same level of granularity as electronic communication. The increasing diffusion of devices with the capability of pervasive and ubiquitous sensing promises to radically change the picture and to allow for addressing face to face communication as efficiently, and with the same granularity, as obtained through electronic communication data.

Social sciences are currently being transformed by these possibilities and Computational Social Science is emerging as a new way to study and predict social behavior [163]. For this new trend to consolidate, common practices, approaches and tools are needed that permit easy capitalization on the results achieved, to quickly circulate new techniques and approaches across the com-

munity and to facilitate comparisons and benchmarking. As in many similar cases, sharable data sets are an essential ingredient of the picture. Data sets and corpora play the indispensable role of permitting extensive comparisons across approaches and techniques and of stimulating and enabling the tackling of new phenomena. This paper contributes to these goals by presenting the SocioMetric Badges Corpus, a new corpus for social interaction studies collected during a 6 weeks continuous period in a research institution and monitoring the activity of 53 people.

The design of the corpus was inspired by the need to provide researchers and practitioners with: a) raw digital trace data that could be used to directly address the task of investigating, reconstructing and predicting people's actual social behavior in complex organizations; b) information about participants' individual characteristics (e.g., personality traits), along with c) data concerning the general social context (e.g., participants' social networks) and the specific situations they find themselves in. Depending on perspective and intent of the researcher, data of type (b) and (c) can help in explaining, predicting and/or discovering relevant behaviors abstracted away from data of type (a), or they can play the role of the investigation target and provide the ground truth for automatic reconstruction of individual traits and states from digital trace data.

In addressing the choice of the wearable device to use, we were guided by the requirement of achieving a multi-scale view of social interactions, from co-presence in a place (squares, buildings, rooms, etc.) to face-to-face proximity of individuals. At present, Bluetooth and Wi-Fi networks allow the collection of data on specific structural and temporal aspects of social interaction [75], offering ways to approximate social interaction as spatial proximity (e.g., through GPS) or as the co-location of wearable devices, e.g., by means of Bluetooth hits [179, 72]. These means, however, do not always yield good proxies to the social interactions occurring between the individuals carrying the devices. Mobile phone traces suffer the same problem: they can be used to model hu-

man mobility [95] with the great advantage of easily scaling up to millions of individuals; they too, however, offer only rough approximations to social interaction in terms of spatial co-location.

For our data collection we resorted to the SocioMetric Badges, wearable sensors able to provide information about: (i) human movement, (ii) prosodic speech features (rather than raw audio), (iii) indoor localization, (iv) proximity to other individuals, and (v) face-to-face interactions [207]. The continuous (accelerometer, audio) and semi-continuous (e.g. infrared, Bluetooth) recording capabilities of the multiple sensors embedded in the SocioMetric Badges[2] make it possible to collect multiplex datasets, spanning over multiple dimensions, and provide a good match for many of the requirements discussed above. Up to now, SocioMetric Badges have been used in several studies to capture face-to-face communication patterns, investigate the relationship among individuals, collective behavior and performance outcomes, such as productivity and job satisfaction [207, 311].

Turning to the personal and contextual information collected, we targeted both stable and transient aspects: (i) *Stable and enduring individual traits (personality, dispositional affectivity and dispositional loneliness)*. These dimensions are expected to be among the (causal) antecedents of people's actual behaviors and therefore to have a good predictive value for them. They were collected by means of standard questionnaires; (ii) *The enduring social ties each individual is involved in and the social networks he/she is part of*. This information was also collected by means of questionnaires; (iii) *The transient states concerning personality, affectivity, creativity and productivity the person goes through in his/her daily life at work*. This information was collected by means of an experience sampling methodology; (iv) *Descriptions of the situations the person was in*. Provided by the person him/herself by means, again, of experience sampling.

[2] www.sociometricsolutions.com

The resulting data set comprises different types of data (digital traces as well as participant-provided assessments of personal traits, states and contextual aspects; data concerning social exchanges through electronic means and data about face-to-face interaction) allowing for a multilayer view of social interaction. It can be used to address many different phenomena ranging from the dynamics of personality and affective states, satisfaction and productivity at work, up to the formation and evolution of social networks, as well as to facilitate the integration of the perspective and tools of social and computational sciences.

4.2.1 Collection Methodology

Recruitment

A total number of fifty-three employees of a research center located in northern Italy were recruited on a voluntary basis to participate in the 6-weeks long study. During introductory meetings, they were provided with detailed information about: the purpose of the study; the data treatment and privacy enforcement strategies adopted; the devices they would be using and the measurements they provide. All participants signed an informed consent form approved by the Ethical Committee of Ca' Foscari University of Venice.

Forty-six out of the fifty-three participants were researchers in computer science belonging to four distinct research groups; the remaining seven participants were part of the full-time IT support staff. Eighty-nine per cent of the participants were male, with a mean age of 36.83 (SD=8.61) years and an average tenure in their current job of 7.48 (SD=6.75) years.
Each participant was assigned a unique ID and all related data were anonymized using these IDs.

Study Design

The experiment was divided into three stages.

Stage 1 Before starting wearing the SocioMetric Badges, participants were given one week to complete an extended initial survey consisting of four sections addressing: (i) personality traits, (ii) dispositional affectivity, (iii) dispositional loneliness, (iv) network-ties at workplace.

Stage 2 During this six week period participants wore the SocioMetric Badges; putting it on at the time they entered the institution's premises and taking it off only when leaving. Issues concerning device maintenance - e.g., battery charging, data downloading, etc. - were taken care of by the study's staff. During this stage, an Experience Sampling Methodology (ESM) was employed to collect information about transient psychological states (personality, affectivity, perceived productivity) and situational aspects. A similar procedure of experience sampling strategy was adopted by [139] and [126] in social psychology studies. Participants completed a short Internet-based survey three times a day. Links to the surveys were automatically administered via email at fixed times and users were granted a temporal window of 2.5 hours to fill the survey before its expiration. Participants were asked to confirm their presence in the institute during the 30 minutes before starting the questionnaire; only if confirmed, their responses would be included in the database.

The experience sampling questionnaire included (i) BIG5 personality scale; (ii) fifteen items concerning affective states, loneliness and two basic emotions (anger and frustration); (iii) a single-item measure of self-perceived creativity and a single-item measure of self-perceived productivity; (iv) five situational items describing the social context. The questions in the experience sampling surveys referred to emotions, behaviors, states, etc., experienced over the 30 minutes prior to the survey.

Stage 3 Participants were asked to complete an extended questionnaire sim-

ilar to the initial one, containing one additional item designed to derive self-perception of face-to-face interactions among participants during the course of the study.

4.2.2 Data Collected: Personal and Situational Data

The measures related to personal and situational factors, as well as to self-perceived social interactions collected through surveys during the three stages of the study are detailed below.

Initial and Final Surveys

The surveys administered during Stage 1 and Stage 3 contained items relating to personality, dispositional affectivity and self-perceived relations with the other participants.

Personality. The Big Five Marker Scale (BFMS) [221] was used to assess personality traits. The BFMS scale is an adjective list composed by 50 items specifically designed to optimize the simplicity of the big-five factor solution in the light of results of psycho-lexical studies on the Italian language [70]. Our sample was composed of 90.56% Italian native speakers; the subjects who were not Italian native speakers received a validated translation of the BFMS.

Dispositional Affectivity. In order to measure dispositional affectivity, a subset of Multidimensional Personality Questionnaire (MPQ) [275] was used with items rated on a 5 point scale from 1= ”strongly disagree” to 5= ”strongly agree”. This sub-scale contained 14 items for Dispositional High Positive Affect such as ’everyday I do some things that are fun’ and ’for me life is a great adventure’ and 17 items for Dispositional High Negative Affect such as ’my feelings are hurt rather easily’ and ’I often lose sleep over my worries’.

Network Ties. To reconstruct self-perceived ties among participants, they were asked to rate to what extent they agreed with the following statements

about each of the other participants: 1) friendship; 2) task related advice; 3) competence; 4) warmth; 5) quality of interaction. A scale from 0 to 7 (0="Not Applicable"; 1="Strongly Disagree" to 7="Strongly Agree") was used.

Experience Sampling

As explained above, a set of short questionnaires were administered three times a day (excluding week-ends) during Stage 2 of the study.

Personality. The ten-item personality inventory TIPI [97] was used to assess personality states. A 7-point scale ranging from 1="Strongly Disagree" to 7="Strongly Agree" was used for responses.

Affect and Loneliness. Respondents were asked to report on a scale from 1 to 5 (1="Very Slightly Or Not At All" and 5="Extremely") to what extent they experienced High Positive Affect (HPA) and/or High Negative Affect (HNA) in the 30 minutes before starting to fill the survey. HPA and HNA were assessed by means of a 6-items shortened version of the Positive and Negative Affect Schedule (PANAS) [300], consisting of 3 items for HPA (α = .78) – "enthusiastic", "interested" and "active" – and 3 items for HNA (α = .83) – nervous, distressed, and upset. Three items were used to measure LPA (α = .70) – "sad", "bored", and "sluggish" – and two items for LNA (α = .77) – "calm" and "relaxed". Finally, two items – "lonely" and "isolated" – were used to measure states of loneliness (α = .87).

Self Perceived Creativity and Productivity. Participants were asked to report their self-perceived productivity (*"how productive were you during the last 30 minutes?"*) and creativity (*"How creative were you during the last 30 minutes?"*)in the previous 30 minutes on a scale from 1="Very Slightly Or Not At All" to 5="Extremely".

Situational Items. Following Fleeson [85], five items were included that described the interactional context of the previous 30 minutes. These items were: 1) *"During the last 30 minutes, how many other people were present*

around you?" to be answered with one of *("0, 1-3, 4-6, 7-9, 10 or more")*; 2) *"I was continuously interacting with the other people around me"*, 3) *"What I was doing was freely chosen by me"*, 4) *"The deadline for what I was doing was very near"*, 5) *"What I was doing was extremely interesting to me"*. The last 4 questions requested answers on a scale from 1="Strongly Disagree" to 5="Strongly Agree".

4.2.3 Data Collected: Digital Data

A large amount of data referring to participants behaviors was collected through the SocioMetric Badges during Stage 2 of the study.

Email and Phone Logs

For each email sent/received to/from another participant in the study, we collected information regarding who sent/received it. Moreover, we recorded the length of the message body, the time, and the list of recipients' IDs. To avoid privacy issues, no information about the content was stored. Phone logs were also collected for each landline telephone of the participants, by recording the phone ID, the timestamp, and the duration of each call. The phone ID was associated to the one or more participants (since in several cases more participants share the same phone). Again, to avoid any privacy issues no information about the content of the phone calls was stored.

SocioMetric Badges Data

SocioMetric Badges are equipped with a microphone, an accelerometer, a Bluetooth sensor and an infrared sensor.

Accelerometer Data. Accelerometer data were sampled with sampling frequency $f_s = 50$ Hz, in a three dimensional space. The accelerometer signal vector magnitude $\mid a \mid$ provides a measurement of the degree of body move-

ment activity by averaging the acceleration's signal power over the three axes. Consistency of body movement is obtained by calculating the standard deviation of the accelerometer signal magnitude for all samples for every 60-seconds interval and subtracting this value from a constant (k = 1) that represents the zero-variation or 100% consistency.

Speech Data. The speech signal was sampled with sampling frequency $f_s = 8$ kHz. A number of basic speech measurements like the signal's amplitude, its standard deviation, minimum and maximum values, mean and variance, were recorded over intervals of 50ms, along with the fundamental frequency and the first 16 mfcc coefficients.

Infrared Data. Infrared (IR) transmissions were used to detect of face-to-face interactions between people. In order for a badge to be detected through IR, two of them must have a direct line of sight and the receiving badge's IR must be within the transmitting badge's IR signal cone of height $h \leq 1$ meter and a radius of $r \leq htan\theta$, where $\theta = \pm 15^o$ degrees; the infrared transmission rate (TR_{ir}) was set to 1Hz.

Bluetooth Data. Bluetooth detection was used as a coarse indicator of proximity between devices. Radio signal strength indicator (RSSI) is a measure of the signal strength between transmitting and receiving devices. The range of RSSI values for the radio transceiver in the badge is (-128, 127). We also used 17 badges as base stations placing them at fixed locations of common interest like the hosting organization's restaurant, the cafeteria and meeting rooms; by detecting participants in close proximity this set up allowed enriching the tracking capability of our study. All SocioMetric Badges, including base stations, broadcast their ID every five seconds using a 2.4 GHz transceiver (TR_{radio} = 12 transmissions per minute).

4.2.4 A Few Statistics

In this section we report quantitative figures and descriptive statistics for the collected data.

Personal and Situational Data

We start by discussing the response rate to the surveys of Stage 1 and 3 and to the experience sampling of Stage 2.

Response Rates **Initial and Final Survey.** All 53 participants took the initial survey (response rate = 100%) while 51 responses were collected in the final survey. This is due to the fact that one participant withdrew from the study during the first week of Stage 2 and another one left the hosting organization during the fourth week. The following discussion considers only the 51 participants sample.

Experience Sampling. We observed that the majority of participants used to leave the organization before 5PM on Fridays afternoon, hence we decided not to consider those surveys in the analysis. To summarize, the 51 participants included in the final sample could complete 14 surveys per week (3 surveys from Mondays to Thursdays plus 2 surveys on Fridays) for a period of 6 weeks, thus 84 surveys in total per participant. Out of the 4284 possible responses (=51 participants * 84 surveys), participants reported not to be at work 536 times (because of meetings outside the institute, participation in conferences, illness, etc.), reducing to 3748 the total number of eligible responses. Out of them, we collected 3147 responses, which is equal to a participation rate of 83,9%. On average, we collected 37,5 responses per signal (SD = 4.2) and 61.7 responses per participant (SD=10.57), over the entire 6-weeks period.

Descriptive Statistics **Initial and Final Survey.** Personality data are normally distributed with kurtosis values falling between -2 and +2, except for Agreeableness in the initial survey. The mood data are not normally distributed as there is a natural skewed pattern to the data: if someone is high positive then it is normal for them to be low negative, as is known to occur with the PANAS scale [57]. Expectedly, the personality and the mood data are very highly correlated between the initial and the final questionnaire with values at least .6 for all measures. This shows that there is consistency in the measures of personality and mood traits of the participants and is confirm by the absence of statistical differences between the the initial and final measure measure for each considered dimension (as tested through ANOVA, $p < .05$).

Experience Sampling. For the transient states concerning personality and affectivity we computed the between-person variance and the within-person variance. More precisely, we define the between-person variance as the variance between each subject's mean and the sample mean, while the within-person variance is based on the difference between each of the subject's score and the subject's mean. With personality states, the within-person variance tends to be higher than the between-person one, a trend that is stronger with Extraversion. These values are in line with findings in the literature on personality states [85] and emphasize the importance of shifting the attention to within-subject variations and their dependence on situational factors when addressing the interplay between personality and actual behavioral manifestations.

SocioMetric Badges Data

We now turn to a few descriptive statistics of the data collected trough SocioMetric Badges. Some problems related to sensor malfunctioning were detected and addressed. In particular, the badges clocks were found to accidentally lose synchronization, probably because of minor accidental shocks they were subject to. This problem was addressed through a data post-processing procedure in

Table 4.13: Statistics for user-to-user infrared face-to-face detection.

	τ =10s	τ =30s	τ =60s
mean	218.09	120.55	84.93
std	562.15	298.87	203.95

which each out-of-sync data slice was re-synced by cross-correlating its clock-time with the one of the most reliable alter detected by means of IR or BT sensors. More specifically, since we adopted a very strict data download procedure (intervals between two subsequent downloads of the same badge were always no longer than 3), we could assign a specific time-window (derived from the recorded download schedule) to all out-of-sync badges. Was a out-of-sync badge found during the data post-processing stage, the records of its IR and BT detected alters were scanned to estimate the faulty badge's clock.

The other problem encountered was related to faulty sensors that would not either work or record data. We could not address this last problem, and we estimate 10% of the data to be missing due to occasional sensor malfunctioning.

Accelerometer and Audio Data. In total we registered 15725.35 hours of bodily activity and 15894.63 of audio data. The difference is explained by the occasional faulty behavior of some sensor leading to missing data.

Infrared Data. Hits from the infrared (IR) sensors were considered as a good proxy for face-to-face interactions; such dyadic interactions were weighted according to their duration. More precisely, the weight w of the face-to-face interaction between subjects X and Y was initialized at 1 at the first hit between their infrared sensors, and incremented of 1 at any subsequent detection happening after at least τ time units from the most recent one. Table 4.13 reports statistics for $\tau = 10, 30, 60$ seconds over the 6 weeks of study.

Bluetooth Data. Bluetooth (BT) hits were used as a proximity measure for a) co-location with other participants, and b) co-location with base stations positioned in strategic places within the premises; BT detections of other ma-

Table 4.14: Statistics for user-to-user and user-to-station bluetooth detection.

user to user	τ =10s	τ =30s	τ =60s
mean	1076.24	653.515	464.93
std	3202.4	1860.78	1284.83
user to base station	τ =10s	τ =30s	τ =60s
mean	321.53	205.18	148.56
std	1377.7	832.75	575.08

chinery (e.g. personal laptops and smartphones) were discarded due to privacy concerns. We assigned a weight w to a tie between two participants or between a participant and a base station; such weight w was then processed following the same strategy detailed above for IR hits. Other approaches are possible, given the raw data in our availability, such as using the RSSI values in order to retain only detections above a threshold. Table 4.14 reports statistics for $\tau = 10, 30, 60$ seconds over the 6 weeks of study.

4.2.5 Discussion and Future Works

In this paper, we have introduced and discussed the design, the methodology and some statistics concerning the SocioMetric Badges Corpus, a new corpus for social interaction studies collected during a 6 weeks contiguous period in a research institution, monitoring the activity of 53 people.

The design of the corpus was inspired by the need to provide researchers and practitioners with: a) raw digital trace data that could be used to directly address the task of investigating, reconstructing and predicting people's actual social behavior in complex organizations; b) information about participants' individual characteristics (e.g., personality traits), along with c) data concerning the general social context (e.g., participants' social networks) and the specific situations they find themselves in.

The resulting data set comprises a rich amount of different types of data: data concerning social exchanges through electronic means and data about face-to-face interaction as well as participant-provided assessments of personal traits, states and contextual aspects. The insights produced by the data remain to be demonstrated by future works. However, the multi-layer nature of the dataset shows the potential for addressing many different research issues.

First of all, the joint collection of stable and enduring individual traits (through the questionnaires administered at Stage 1 and Stage 3), of transient states concerning subjects' personality (by means of the experience sampling of Stage 2) and of the raw data, provides the means to shift the focus from the prediction of personality traits right away from actual behavioral manifestations as in [166, 25, 256] to their reconstruction on the basis of the distribution of transients personality states [84]. In socio-psychological literature, a personality state is a specific behavioral episode wherein a person behaves as more or less introvert/extravert, neurotic or open to experience, etc. This approach could be beneficial not only for the automatic computation of personality, but also for the long-term goal of predicting/explaining individual behavior from individual characteristics. Recently, some research efforts [255] have also tackled the explanation of the mutual influence between individual characteristics and network structure. Our dataset gives also the chance to trace the origin of network structure to its elemental relational and affective components: the ongoing micro-interactions between people in organizations, and the specific emotions they experience during those interactions. With existing research having demonstrated the importance of affect in network formation and individual performance, our data can be used to uncover the finer micro-structure of affect in organizations.

Which interactions elicit different discrete emotions? How do these emotions shape subsequent interactions over time? And how do these ongoing micro-interactions result in aggregate network structure? Moreover, our dataset

gives also the opportunity of testing how the spreading of emotions inside an organizations can influence performance, creativity and job satisfactions of the employees.

Finally, the multi-layer longitudinal nature of the corpus described in this paper makes it highly valuable for network-based studies on organizational behavior.

4.3 Ego-Centric Graphlets for Personality and Affective States Recognition

Affect and personality permeate people's daily and working life and also the interdependent relationships they usually hold with bosses, colleagues, and subordinates. Several studies showed the relationships between personality and e.g. job performance [22] and job satisfaction [138].

At the same time, an affective revolution has taken place, in which academics and managers have begun to appreciate how an organizational approach that integrates employee affect provides a more complete perspective [23]. Previous studies outlined effects of affect on performance [259], decision making [129], and prosocial behavior [241].

Usually, we can think of affect and emotions both as *states* or *traits*. These two levels differ in terms of the extent to which they are deeply characteristic of the individual, and therefore the extent to which they are mutable or immutable. At the same time, traditionally scientific psychology has developed a view of personality as a higher-level abstraction encompassing traits, sets of stable dispositions towards action, belief and attitude formation.

The problem with this approach to personality is that it assumes a direct and stable relationship between being extravert and acting extravertedly (e.g., speaking loudly, being talkative, etc.). Extraverts, on the contrary, can often be silent and reflexive and not talkative at all, while introverts can at time exhibit

extraverted behaviors.

While personality studies have often dismissed these fluctuations of actual behavior as statistical noise, it has been suggested by Fleeson [84, 85] that they can give a valuable contribution to personality prediction and to the understanding of the personality/behavior relationship. The social psychology literature has recently coined the term *personality states* to refer to concrete behaviors that can be described as having a similar content to the corresponding personality traits. In other words, a personality state describes a specific behavioral episode wherein a person behaves more or less introvertly/extravertly, more or less neurotically, etc.

In this work, we investigate the influence played by specific situational factors, the face-to-face interactions and the proximity interactions with alters, over the ego's expression of a particular affective/emotional state or a specific personality state in a work environment. In particular, how the details and the complexity of the social network structure of the interacting alters can play a significant role in predicting the affective and personality states of the ego. To this end, we represent people's interactions as *graphlets*, induced subgraphs representing specific patterns of interaction, and design classification experiments with the target of predicting the subjects' self-reported personality and affective states.

Being able to capture the local structure of interactions, graphlets represent a promising methodology to study interactions between humans in the online and offline worlds.

Graphlets have extensively been employed to study properties of biological networks, e.g. to discover invariant patterns characterizing specific properties of enzymes and small molecules. Being able to capture the local structure of interactions, graphlets represent a promising methodology to study interactions between humans in the online and offline worlds.

We investigate graphlets centered on the reference node (the *ego*), embed-

ding information on the state of the alters and their interactions in order to recognize the affective/personality state of the ego. We explore how interaction patterns, encoded as graphlets, gathered from two distinct sensory channels, Bluetooth (BT) and infrared (IR), affect recognition of personality and affective states.

4.3.1 Related Works

Several studies in social psychology have revealed links between positive affect and social activity. Recently, Hatfield et al. coined the term *emotional contagion* [115] to describe the process by which people "catch" emotions from each other. Positive and negative moods also spread during long periods [88] and over workplace interactions [24].

Inspired by the susceptible-infected-susceptible (SIS) disease model, Hill et al. proposed a mathematical model for the contagion of long-lasting emotional states in a self-reported social network [117]. In social and ubiquitous computing, researchers have explored the associations between mood and social interactions captured by mobile phones [198]. These studies assume that for detecting or predicting if an individual is in a positive or negative affect state it is enough to look at the number of individuals with whom he or she is in contact, and possibly at their state. Instead, we investigate how the *structure* of the interaction network can play a significant role in predicting the affective and personality states of the ego.

Regarding personality, researchers have started exploring the wealth of behavioral data made available by cameras and microphones in the environment [222, 167], smartphones [255, 65], wearable sensors [205] in order to automatically classify personality traits.

However, the general approach of all these previous works is to isolate promising correlates of the targeted traits for classification or regression. All these works adopted the so-called person-perspective on personality and target per-

sonality traits prediction or classification and not personality states prediction or classification.

4.3.2 Dataset

For this study we exploited the SocioMetric Badges Corpus [165], described in Section 4.2, a multimodal corpus designed to capture the psychological and situational aspects of the daily lives of employees in an organizational structure. The data were collected in a research institute for six weeks on a sample of 54 subjects during their working hours. Males predominated (90.8%) while the average age was 36.83 ± 8.61 years. The data were collected using wearable sensors called Sociometric Badges.

Table 4.15 reports basic dataset statistics on the target behavioral states.

Table 4.15: Mean and median daily values for personality and affective states in the dataset used.

state	mean	median
Extroversion	4.07	4.0
Agreeableness	5.13	5.5
Conscentiousness	5.53	6.0
Emotional Stability	5.54	6.0
Openness/Creativity	4.50	4.5
High Positive Affection	3.12	3.3
High Negative Affection	1.42	1.3

In this work, we exploit information from the infrared (IR) and Bluetooth (BT) sensors.

4.3.3 Graphlet-based Approach

We define a binary classification task for each subject and each personality and affective state. This is done by mapping the state of a given subject at a certain deadline from $\{1, \ldots, 7\}$ to $\{0, 1\}$ using its median value for the subject as a threshold. Therefore, negative labels represent cases where the subject was found below his/her median. One of the main contributions of this paper lies in the encoding of the subjects' interactions as *graphlets*, defined as induced subgraphs of a larger network, providing a succinct representation of social structure. In the Bioinformatics and Computational Biology domains, graphlets have been introduced for the study of large biological networks, for e.g. network alignment [225]. Recently, graphlet analysis has been applied to Facebook messaging and historical crime data [253].

We investigate their effectiveness in the context of a human interaction network, for the prediction of behavioral determinants such as personality and affective states.

Starting from the network of interactions between subjects, we extract for each subject the graphlets representing his/her local interactions. In this work, we consider all possible graphlets up to 4 nodes, as shown in Figure 4.2, where the double circle represents the reference subject and his/her interacting partners can have multiple patterns of reciprocal interaction. Furthermore, the graphlets embed information on the current (binary) state of the alters (but not of the reference subject whose state is to be predicted), in order to account for possible influence and propagation effects.

For each deadline, we extract graphlet-based features from sensory data gathered over the previous 3 hours. We discretize each 3-hour window in 15-minutes slices in order to represent the evolution of the interaction patterns over time, taking into consideration the neighbours' states in order to account for situational influence effects. To do so, we count occurrences of graphlet con-

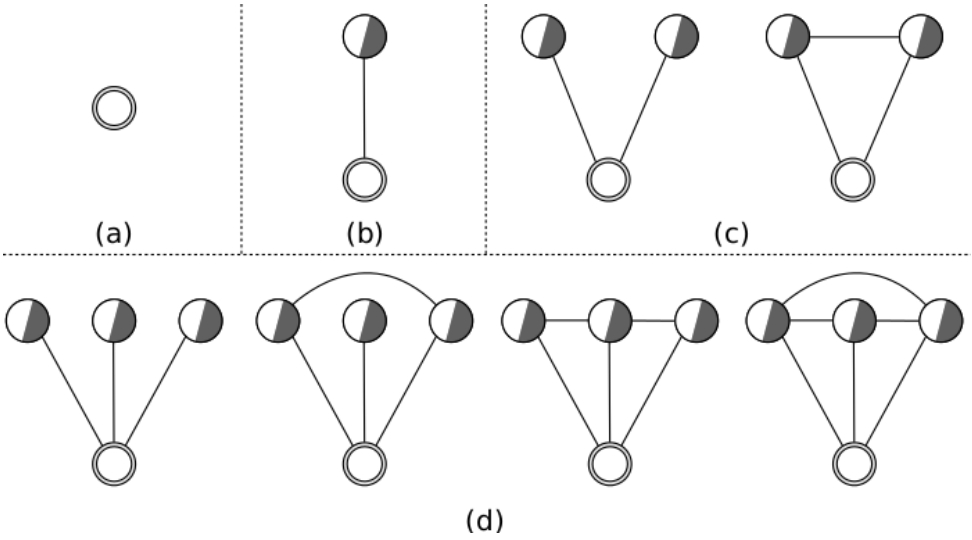


Figure 4.2: Graphlet configurations used. Bottom nodes (double circled) represent the reference subject (the ego), while top nodes represent alters and their binary state.

figurations and build a histogram; then, we average the histograms obtained for each slice and obtain a feature vector representative of the 3-hours window under analysis. Finally, we use the latter to predict the ego's state at the deadline under analysis.

In our setup, two kinds of missing data are possible: i) missing labels (i.e. surveys not filled by the subjects); ii) missing interactions, in which case the interaction graphs will be empty. In both cases we exclude the deadline under analysis from the training and testing stages.

4.3.4 Experimental Setup

To understand the influence of alters on a given subject, we predict its state based on features derived from the labeled graphlets.

We build a linear Support Vector Machine (SVM) model [53] for each agent and each target state, evaluating its performance in a leave-1-week-out cross-validation procedure. We employed the LibLinear [82] library with ℓ_1 regularization, which tends to produce sparse models (with few non-zero weights).

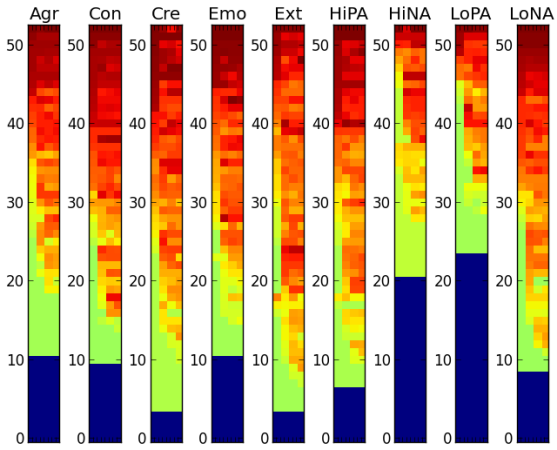


Figure 4.3: Performance obtained using infrared sensory data.

The learned models prioritize informative features, leading to robust handling of noise, and are simpler to interpret. To avoid any bias in the interpretation of the results, we discard all agent/target state pairs for which one class (i.e. positive or negative) covers at least 75% of the instances. This occurs when subjects exhibit very little variance on the labels (see Low PA - High NA in Figure 4.3), and thus many instances fall on the median value itself. We build models of increasing complexity by considering graphlet-based features made of up to one, two and three alters respectively, and evaluate the performance of each model in order to assess the predictive power of different levels of interaction. We compare these models against each other and against a majority classifier (i.e. a classifier that always predicts the class with more instances in the training set), which we use as a baseline.

4.3.5 Experimental Results

Figure 4.3 shows the performances obtained using features extracted from IR sensors (those for BT are similar). Each row represents a subject in our dataset. Each bar represents a target state. The left-most columns within each target display the performance of our baseline (a majority classifier), while subsequent

ones represent those of the SVM models built using graphlets of increasing complexity. Each pixel represents the average f_1 measure [3] over all folds for a specific subject/target pair. The values range from low (light green) to high (red). Blue values represent cases with missing labels or highly unbalanced classes, which we ignore in the following analysis.

For the sake of readability, we cluster the subjects (rows) by performance, using k-means with $k = 3$, and plot them from worst (bottom) to best (top). The plot shows that the graphlets are very predictive in the middle two clusters, where the performances tend to transition from bad (green) to good (red). This trend is not as clear on the high-performance cluster (top). The latter represents cases with an unbalanced, overly-positive data component (yet no more than 75% of the total), where the agents show little state variance and interactions can not be useful.

To better understand the above trends we compute Win/Loss matrices for each target state and cluster: for each pair of feature-sets and for all subjects in a cluster we count the number of times a feature-set outperforms the others. Table 4.16 reports the resulting matrices for all target states and clusters for IR (results for BT are similar). Intuitively, positive values below the diagonal imply that the more complex feature-sets are informative and bring performance benefits; positive values above the diagonal have the opposite meaning. Analysis of the matrices shows that using more complex graphlets constantly improves the performances in the low and middle clusters. The behavior on the high-performance cluster is less clear, due to the aforementioned unbalance in the agent states distribution.

We then compare the results of different target states for the two clusters (low- and middle-) where performances improve. For each matrix we compute the percentage of cases in which higher degree graphlets outperform (underperform) lower-degree ones, using the normalized sum of all elements above

[3] The f_1 measure is defined as the harmonic mean of precision and recall.

Table 4.16: Win/Loss matrices for the 3 performance clusters for the different states predicted using IR sensors.

State	Low-perf. cluster				Mid-perf. cluster				High-perf. cluster			
Agreeableness	0	0	0	0	0	1	1	1	0	6	7	8
	1	0	0	0	13	0	4	3	7	0	6	8
	3	3	0	2	14	10	0	6	6	5	0	6
	3	3	0	0	14	10	5	0	5	3	5	0
Conscientiousness	0	0	0	0	0	0	0	0	0	8	9	10
	1	0	0	1	14	0	2	1	9	0	6	7
	1	1	0	1	15	12	0	6	9	8	0	5
	2	1	1	0	15	13	4	0	8	6	5	0
Creativity	0	0	0	0	0	1	1	1	0	13	11	11
	4	0	2	2	15	0	6	8	3	0	8	6
	5	3	0	0	15	9	0	3	6	7	0	5
	7	5	6	0	15	7	5	0	6	8	4	0
Emotional Stability	0	0	0	0	0	2	2	2	0	3	4	3
	7	0	2	2	16	0	6	6	3	0	4	3
	9	5	0	1	16	11	0	6	2	3	0	3
	9	6	4	0	16	11	4	0	3	5	2	0
Extraversion	0	2	1	1	0	0	1	0	0	8	9	8
	7	0	0	2	20	0	9	9	4	0	5	7
	9	11	0	5	19	9	0	5	4	6	0	3
	10	10	3	0	20	9	5	0	5	4	5	0
High PA	0	1	2	2	0	5	3	4	0	10	10	9
	5	0	2	2	15	0	4	5	1	0	5	6
	7	6	0	3	17	14	0	4	1	4	0	2
	7	6	2	0	16	13	6	0	2	3	4	0
High NA	0	0	0	1	0	1	0	0	0	2	0	0
	9	0	2	2	10	0	3	2	0	0	0	0
	9	6	0	3	12	8	0	3	1	2	0	0
	8	7	3	0	12	8	4	0	1	2	0	0
Low PA	0	0	0	0	0	0	0	0	0	4	3	3
	3	0	2	1	11	0	5	4	3	0	2	3
	3	2	0	0	13	7	0	2	4	4	0	2
	4	3	2	0	13	8	7	0	4	3	1	0
Low NA	0	0	0	0	0	0	0	0	0	11	7	9
	6	0	2	3	11	0	7	6	3	0	4	5
	8	4	0	2	12	4	0	4	7	8	0	5
	9	5	4	0	12	5	2	0	5	7	4	0

(below) the diagonal. Table 4.17 lists the difference between the two values, i.e., the relative percentage of cases for which more complex graphlets are beneficial.

The results confirm that graphlet-based features are predictive of personality and affective states: all values are positive and well above the mean of a random classifier (i.e. 0).

Finally, we sort such values for IR and BT, and compute the Spearman rank correlation between the two lists. The correlation coefficient is found to be -0.4 (p-value$= 0.28$), and indicates that the two channels can be effectively exploited for different target states, and support the intuition that the two channels capture different behavioral manifestations: BT captures proximity in a broad-cast manner (i.e. many-to-one), while IR is restricted to face-to-face (one-to-one) interactions.

Table 4.17: Performance improvement factors, in ascending order. Value range is $[-1, 1]$.

IR		BT	
Target	Improvement	Target	Improvement
High PA	0.51	Creativity	0.44
Low NA	0.55	Conscientiousness	0.55
Extraversion	0.58	Emotional Stability	0.62
Emotional Stability	0.59	High PA	0.63
Creativity	0.6	Agreeableness	0.67
Agreeableness	0.63	Low PA	0.67
Low PA	0.69	Extraversion	0.69
High NA	0.70	Low NA	0.71
Conscientiousness	0.76	High NA	0.81

4.3.6 Conclusion

In this work we investigated new perspectives on affect and personality states recognition, studying in particular the influence on the ego's state of face-to-face

and proximity interactions with alters. To this end, we propose a graphlet representation of the ego-network, computed using two distinct sensory channels (Bluetooth and infrared), to predict the ego's state. The advantage of graphlets over other representations is that they capture not only the number of interactions, but also their structure at different levels of complexity.

Our results demonstrate that the graphlet-based representation consistently contributes to recognition improvements over a baseline. Furthermore the amount of improvement tends to increase with graphlet complexity. These results show the feasibility of the proposed approach, and hopefully encourage further research.

We also find that distinct sensory channels play different roles for distinct target states: e.g. complex graphlets derived from IR have a large impact for Conscentiousness, while those derived from BT do not. The opposite trend is observed for Low NA. These findings support the intuition that the two channels capture different concrete behaviors: BT reflects proximity in a broadcast manner (i.e. many-to-one), IR is restricted to face-to-face (one-to-one) interactions.

4.4 A Human-Centric Study on the Economics of Personal Mobile Data

The number of mobile phones actively in use worldwide today is about 5 billion, with millions of new subscribers every day[4]. Mobile phones allow for unobtrusive and cost-effective access to previously inaccessible sources of behavioral data such as location, communications (calls and text messages), photos, videos, apps and Internet access [157]. Hence, a result of the ever-increasing adoption of these devices is the availability of large amounts of *personal data* related to habits, routines, social interactions and interests [157, 179].

However, the ubiquitous collection of personal data raises unprecedented

[4]http://www.ericsson.com/ericsson-mobility-report

privacy challenges. Users typically have to make decisions concerning the disclosure of their personal information on the basis of a difficult tradeoff between data protection and the advantages stemming from data sharing. Perhaps more importantly, people are typically not involved in the life-cycle of their own personal data – as it is collected by websites and mobile phone apps, which results in a lack of understanding of who uses their data and for what.

Several researchers have proposed and investigated new user-centric models for personal data management, which enable individuals to have more control of their own data's life-cycle [217]. To this end, researchers and companies are developing repositories which implement medium-grained access control to different kinds of personally identifiable information (PII), such as *e.g.* passwords, social security numbers and health info [297], and more recently location [66, 118, 199] and personal data collected online by means of smartphones or wearable devices [66, 288].

Previous work has introduced the concept of *personal data markets* in which individuals sell their own personal data to entities interested in buying it [5]. Buyers are likely to be companies and researchers, while sellers are individuals who receive compensation for sharing their own data. Riederer *et al.* [237] have recently proposed a mechanism called *transactional* privacy, devised to maximize both the user's control of their own PII and the utility of a data-driven market.

In the context of prospective personal data markets that offer increased transparency and control, it is of great importance to understand the value that users put to their own PII. Recently, Carrascal *et al.* [39] used a refined Experience Sampling Method (rESM) [44] and a reverse second price auction to assess the monetary value that people assign to their PII shared online via websites – *e.g.* keywords used in a search engine, photos shared in a social network, etc. However, the authors focus only on web-browsing behaviors without taking into account behaviors and personal information that can be captured by

mobile phones.

Taking Carrascal *et al.* [39] as an inspiration, in this paper we investigate the monetary value that people assign to different kinds of PII as collected by their mobile phone, including location and communication information.

We carried out a comprehensive 6-week long study in a living lab environment with 60 participants and adopted a Day Reconstruction Method [141] along with a reverse second price auction mechanism in order to poll and collect honest monetary valuations from our sample.

The main contributions of this paper are:

1. Quantitative valuations of mobile PII as collected by a 6-week long study conducted in the wild;

2. Qualitative feedback on the valuations provided by each participant as gathered by an End of Study (EoS) survey;

3. A segmentation of PII valuations and findings based on 4 categories of mobile PII (communications, location, media and apps), 3 levels of complexity (individual, aggregated, processed), and one level of temporal granularity (daily);

4. A set of key insights about people's sensitivities and valuations of mobile PII and implications for the design of mobile services that leverage mobile PII.

4.4.1 Related Work

In recent years, researchers have analyzed the factors that can influence a person's disclosure behavior and economic valuation of personal information. Demographic characteristics, such as gender and age, have been found to affect disclosure attitudes and behavior. Several studies have identified gender differences concerning privacy concerns and consequent information disclosure

behaviors: for example, women are generally more protective of their online privacy [86, 119]. Age also plays a role in information disclosure behaviors. In a study on Facebook usage, Christofides *et al.* [48] found that adolescents disclose more information.

Prior work has also emphasized the role of an individual's stable psychological attributes - *e.g.* personality traits - to explain information disclosure behavior. Korzaan *et al.* [155] explored the role of the Big5 personality traits [54] and found that Agreeableness – defined as being sympathetic, straightforward and selfless, has a significant influence on individual concerns for information privacy. Junglas *et al.* [140] and Amichai-Hamburger and Vinitzky [10] also used the Big5 personality traits and found that Agreeableness, Conscientiousness, and Openness affect a person's concerns for privacy. However, other studies targeting the influence of personality traits did not find significant correlations [245]. More recently, Quercia *et al.* [229] found weak correlations among Openness to Experience and, to a lesser extent, Extraversion and the disclosure attitudes on Facebook. In 2010, Lo [176] suggested that Locus of Control [239] could affect an individual's perception of risk when disclosing personal information: internals are more likely than externals to feel that they can control the risk of becoming privacy victims, hence they are more willing to disclose their personal information [315].

Individual differences are also found when providing economic valuations of personal data [4, 39]. For instance, some individuals may not be concerned about privacy and would allow access to their data in exchange for a few cents, whereas others may only consent if well paid. Recently, Aperjis and Huberman [11] proposed to introduce a realistic market for personal data that pays individuals for their data while taking into account their own privacy and risk attitudes.

Previous research has shown that disclosure [152] and valuation [61, 120] depend on the kind of information to be released. Huberman *et al.* [120] re-

ported that the valuation of some types of personal information, such as the subject's weight and the subject's age depends on the desirability of these types of information in a social context. Some empirical studies have attempted to quantify subjective privacy valuations of personal information in different contexts, such as personal information revealed online [113], access to location data [59], or removal from marketers' call lists [291].

These studies can be classified into two groups. The first and larger group includes studies that explicitly or implicitly measure the amount of money or benefit that a person considers to be enough to share her/his personal data, namely their *willingness to accept* (WTA) giving away his/her own data (see for example [59, 121]). The second and smaller group includes studies about tangible prices or intangible costs consumers are *willing to pay* (WTP) to protect their privacy (see for example, [3, 281]). In our paper, we do not deal with WTA vs WTP, but we focus on WTA for PII captured by mobile phones (communications, apps and media usage, locations).

A growing body of studies in the fields of ubiquitous and pervasive computing and human-computer interaction focuses on location sharing behavior and has highlighted the role played by the recipient of sharing (who can access the information), the purpose, the context, how the information is going to be used [21, 52, 174, 278, 308] and the level of granularity of the information shared [173].

Finally, studies have suggested the importance of analyzing people's actual behavior rather than attitudes expressed through questionnaires because often the actual behavior of people deviates from what they state [133].

Building upon previous work, in this paper we investigate the monetary value that people assign to different kinds of PII as collected by their mobile phone, including location and communication patterns. In particular, we carry out a comprehensive 6-week long study in a living lab environment with 60 participants and adopt a Day Reconstruction Method [141] and a reverse second price

auction mechanism in order to poll and collect honest monetary valuations from our sample.

4.4.2 Methodology

Next, we describe the methodology followed during our 6-week study.

The Living Laboratory

The Living Laboratory where we carried out our study was launched in November of 2012 and it is a joint effort between industrial and academic research institutions. It consists of a group of more than 100 volunteers who carry an instrumented smartphone in exchange for a monthly credit bonus of voice, SMS and data access. The sensing system installed on the smartphones is based on the FunF[5] framework [6] and logs communication events, location, apps usage and photos shot. In addition, the members of the living lab participate in user-studies carried out by researchers.

The goals of this living lab is to foster research on real-life behavioral analysis obtained by means of mobile devices, and to deploy and test prototype applications in a real-life scenario. One of the most important features of such a lab is its ecological validity, given that the participants' behaviors and attitudes are sensed in the real world, as people live their everyday life, and not under artificial laboratory conditions.

All volunteers were recruited within the target group of young families with children, using a snowball sampling approach where existing study subjects recruit future subjects from among their acquaintances [96]. Upon agreeing to the terms of participation, the volunteers grant researchers legal access to their behavioral data as it is collected by their smartphones. Volunteers retain full rights over their personal data such that they can order deletion of personal

[5] http://funf.org

information from the secure storage servers. Moreover, participants have the choice to participate or not in a given study.

Upon joining the living lab, each participant fills out an initial questionnaire which collects their demographics, individual traits and dispositions (*e.g.* Big Five personality traits, trust disposition, Locus of Control, etc.) information.

Participants

A total of 60 volunteers from the living lab chose to participate in our mobile personal data monetization study. Participants' age ranged from 28 to 44 years old ($\mu = 38$, $\sigma = 3.4$). They held a variety of occupations and education levels, ranging from high school diplomas to PhD degrees. All were savvy Android users who had used the smartphones provided by the living lab since November 2012. Regarding their socio-economic status, the average personal net income amounted to €21169 per year ($\sigma = 5955$); while the average family net income amounted to € 36915 per year ($\sigma = 10961$). All participants lived in Italy and the vast majority were of Italian nationality.

Procedure

Our study ran for six weeks from October 28th, 2013 to December 11th, 2013. At the beginning of the study, participants were explained that the study consisted of three phases:

1. An initial questionnaire, which focused on their general perception of privacy and personal data;
2. A daily data collection phase that lasted 6 weeks where participants answered daily surveys to valuate their mobile personal data;
3. A final survey that aimed to clarify the results obtained and to collect qualitative feedback from participants.

Daily Surveys Ad-hoc `java` code was developed and scheduled to run on a secure server each night in order to automatically generate personalized daily surveys for each participant. The survey questions were generated based on the mobile data collected during the previous day. Everyday, at 12PM, participants received an SMS reminding them to fill out their survey via a personalized URL (through a unique hash).

In order to test the live system and identify bugs, we ran a pilot for 10 days with a small set of volunteers who were not participants in the study. In addition, we allocated an additional *training* week prior to starting the actual study so participants would get accustomed to the survey/auction scheme.

The content of the daily surveys is described next.

4.4.3 Collected Data

Next we describe the data that we collected during the study.

Mobile Personal Data

We collected 4 categories of mobile personal data:

1. *communications*, in the form of calls made/received;
2. *locations*, collected by the device GPS sensor every ~ 5 minutes;
3. *running applications*, sampled every 25 minutes;
4. *media*, number and timestamp of pictures taken and obtained by monitoring the device file system.

The sampling rates for the different categories of data were empirically determined in order to have good resolution without significantly impacting the device's battery life.

We probed participants about three levels of complexity for each category of data:

Data Type	Individual	Aggregated	Processed
Communications	A call event*	# of calls or diversity	Total duration of calls
Location	A place visited*	# of places visited	Total distance covered
Running Apps	An app running*	# of apps running	An app running for N minutes in the**
Media	A picture shot*	# of pictures shot	Pictures shot in the**

Table 4.18: Categories of personal data probed in the surveys. Include [*: at time hh:mm; **: night (12AM-6AM),morning (6AM-12PM), afternoon (12PM-18PM), evening (18PM-12AM)]. Questions referred to data collected the previous day.

1. *individual*, encompassing individual data points (*e.g.* a call made/received, a picture taken, a specific GPS location);

2. *aggregated*, portraying cumulative event information (*e.g.* number of places visited, number of calls made/received);

3. *processed*, depicting higher level information derived from aggregated data and time (*e.g.* a given application has been running for N minutes, distance travelled).

For each data category and level of complexity, participants were asked to fill out daily surveys that asked them about data from the previous day for each category and for a specific level of complexity (up to 4 questions per day). For each question in the surveys, participants always had the option to opt-out and not sell that particular piece of information.

Next, we describe in detail the 4 categories and the 3 levels of complexity of mobile personal data that we collected in this study, which are summarized in Table 4.18.

Communications *Individual* communication data was restricted to voice calls made/received; missed calls were discarded. With respect to *aggregated* communications data, we alternated between two different aggregated variables on a weekly basis: on even weeks subjects were asked to monetize information about the total number of calls made/received during the previous day, while

on odd weeks they were asked about call diversity, *i.e.* the number of different people that they talked to on the phone during the previous day. Examples of questions related to aggregate communications are `"Yesterday, you made/received 8 phone calls"`, or `"Yesterday, you spoke on the phone with 3 different persons"`. The *processed* communication variable referred to the total duration of calls in the previous day, resulting in questions such as `"Yesterday, you spoke on the phone for a total of 52 minutes"`.

Location *Individual* location referred to a specific place visited by the participant in the previous day. Semantic information associated to GPS locations was derived via reverse geo-coding using Yahoo Query Language. For individual locations, details on street, neighborhood and town were included in the question. For example, `"Yesterday, at 23:56 you were in Via Degli Orbi 4, Trento"`.

Location data was spatially clustered over the reference time-range using a threshold of 100 meters to generate the *aggregated* location question (*e.g.* `"Yesterday you have been in 23 different places".`). Finally, the *processed* location variable referred to the total distance traveled in the previous day, resulting in questions such as `"Yesterday you covered a total distance of 13km"`.

Running Applications With respect to running apps, the *individual* variable included the timestamp and the name of the app running in the foreground. *Aggregated* apps referred to the total number of different apps that the participant ran the previous day, whereas *processed* apps referred to the total number of minutes that a particular app was running over a specific time in the previous day.

Examples of app-related information questions for each level of complex-

ity are, "Yesterday, at 10:23 you were using the Firefox Browser application", "Yesterday 9 applications were running on your device", and "Yesterday night, the Facebook applicati run on your device for 82 minutes", respectively.

Media *Individual* media asked participants about the fact that they shot a photo at a specific time ("Yesterday, at 14:23, you shot one picture"). For legal privacy reasons, the questions referring to individual media data could not include the actual picture they referred to. *Aggregated* media referred the total number of pictures shot the previous day (*e.g.*, "Yesterday you took 9 pictures"). Finally, *processed* media probed participants about their photo-taking activity during specific times of the day (*e.g.*, "Yesterday morning you took 4 pictures").

Individual Traits Data

As previously mentioned, upon joining the lab each participant filled out 4 questionnaires to collect information about their personality, locus of control, dispositional trust and self-disclosure behaviors.

The Big Five personality traits were measured by means of the BFMS [221] questionnaire, which is validated for the Italian language and covers the traditional dimensions of Extraversion, Neuroticism, Agreeableness, Conscientiousness, and Openness [54].

Participants also provided information about their *Locus of Control* (LoC) [239], a psychological construct measuring whether causal attribution for subject behavior or beliefs is made to oneself or to external events and circumstances. The LoC measures whether the outcomes of a set of beliefs are dependent upon what the subject does (internal orientation) or upon events outside of her/his control (external orientation). LoC was measured by the Italian version of Craig's Locus of Control scale [83].

Moreover, we collected information about the participants' *dispositional trust*. Rotter [239] was among the first to discuss trust as a form of personality trait, defining interpersonal trust as a generalized expectancy that the words or promises of others can be relied on. In our study, we resort to Mayer and Davis's Trust Propensity Scale [187].

Finally, we targeted the *self-disclosure* attitudes of our subjects. Self-disclosure has been defined as any message about the self that an individual communicates to another one [56, 304]. We used Wheeless' scale [304] measuring five dimensions of emphself disclosure, namely (i) amount of disclosure, (ii) positive-negative nature of disclosure, (iii) consciously intended disclosure, (iv) honesty and accuracy of disclosure, and (v) general depth or intimacy of disclosure. Wheeless' scale has been utilized to measure self-disclosure in online communication and in interpersonal relationships [304].

Auctions of mobile PII

The personalized daily survey asked each participant to place a bid to sell one piece of their mobile personal information for each of the four categories of study (communications, location, apps and media), for a specific level of complexity (individual, aggregated or processed) and for the previous day. The winner of each auction won the monetary value associated with that auction. In exchange, (s)he sold that particular piece of information to the Living Lab which could use it for whatever purpose it wanted.

In order to ensure a balanced sample, surveys were generated by rotating the different levels of complexity described above, such that each day participants placed bids in up to 4 auctions: one for each category of personal information and for a particular level of complexity (individual, aggregated or processed). Note that in the case a participant did not generate any data for a particular category, s(he) was still asked to provide a valuation to the fact that there was no data in that category, *e.g.* "`Yesterday you did not make`

`any phone call"`.

The participants' bids entered a reverse second-price auction strategy, *i.e.*, the winner was the participant(s) who bid the lowest, and the prize was the second lowest bid. The choice of this auction mechanism was due to the following reasons: (1) the mechanism is truth telling given that the best strategy for the auction participants is to be honest about their valuation [188], (2) it is easy to explain and understand, and (3) it has sucessfully been used before to evaluate location information in [61] and Web-browsing information in [39].

Question	**mean**	**st_dev**
Q1. I am concerned about the protection of the data collected by my smartphone	4.7	1.6
Q2. I trust the applications I install and run on my smartphone wrt how they use my data	3.7	1.5
Q3. I trust telco providers with respect to how they use my data	3.4	1.4
Q4. I always read the privacy terms and conditions for the applications I use	2.7	1.6
Q5. I know the legislation on mobile communication data protection	2.5	1.5

Table 4.19: Questions asked in the Initial questionnaire, and responses statistics. The 7-point likert scale used goes from *1-Totally Disagree* to *7-Totally Agree*.

Interventions, *i.e.* individual communications of auction outcomes to participants, took the form of e-mails sent every Thursday.

In order to evaluate possible effects of winning frequency on bidding behavior, we employed two different auction strategies for the first and second halves of the study. During the first 3 weeks (phase 1), we carried out weekly auctions on Wednesday, taking into account all bids that had been entered during the previous 7 days for each category. Therefore, in this phase, 12 weekly auctions took place with the daily bids for each category and level of complexity (4 categories x 3 levels of complexity). During the last 3 weeks of the study (phase 2), we switched to daily auctions, resulting in a total of 12 auctions per day. In addition, the sample of bidding participants was split into 3 random subsets in order to increase their chances of winning.

Email interventions were always on Thursdays and therefore this change was transparent to participants. Interventions were sent to all participants, whether they had won auctions or not. In the case of winners, the intervention email included the specific piece of information that the participant had sold, the corresponding winning bid, and the amount won. In the case of losers, the intervention email simply communicated the participant that s(he) did not win any of their auctions. All emails were kept neutral for both winners and losers.

In total, 596 auctions were run during the entire study (36 in the first three weeks, 560 afterwards).

Pre- and Post- Study Questionnaires

As previously explained, at the beginning and at the end of the data collection participants were required to fill out initial and end-of-study (EoS) questionnaires. The initial questionnaire consisted of 5 questions (see Table 4.19) and was used to gather information about the participants' perception of privacy issues related to mobile personal data. From the responses provided to this survey, we notice that participants are concerned about mobile PII protection (Q1) but do not tend to read the Terms of Service (Q4) nor are aware of current legislation on data protection (Q5). Moreover, they do not seem to trust how neither application providers (Q2) nor telecom operators (Q3) use their data.

The EoS survey was designed to gather additional quantitative and qualitative information from our participants after the data collection was complete. In particular, we asked participants to put a value (under the same auction game constraints) on category-specific *bulk information – i.e.* all the data gathered in the study for each category. For instance, in the case of location information, a visualization of a participant's mobility data collected over the 6-weeks period was shown in the Web questionnaire (as depicted in Figure 4.4) and the participant was asked to assign it a monetary value. Furthermore, for each category, we asked participants about the minimum/maximum valuations given during

the study, in order to understand the reasons why they gave these valuations. Table 4.20 contains all the questions of the EoS survey.

The EoS questionnaire was administered through a slightly modified version of the same Web application used for the daily surveys. The main difference are the visualizations of the collected data.

Question	**Type**
Q1. This {map‖chart} shows the information about {locations‖communications‖apps‖media} we collected during this study. What is the minimum amount of money you would accept to sell it in anonymized/aggregated form?	numeric
Q2. On day {dd/MM} you assigned a value of {*min-bid per category*} to the information [{*least valued info per category*}]. This was your minimum bid. Why?	multi-choice*
Q3. On day {dd/MM} you assigned a value of {*max-bid per category*} to the information [{*most valued info per category*}]. This was your maximum bid. Why?	multi-choice*
Q4. Imagine there was a market in which you could sell your personal information (e.g. information about people you called, places you've been, applications you've used, songs you've listened to, etc.). Who would you trust to handle your information? Please, order the following entities from most to least trusted.	rank**
Q5. The category {locations‖communications‖apps‖media} is the one that you refused to sell the most ({*percentage of opt-outs*}). Why?	free-text

Table 4.20: Questions asked in the EoS questionnaire.
*included: *Fair value*, *Test/Mistake*, *Other (free text)*. For minimum-bid related questions additional options were *To win the auction*, *Info not important*; conversely, for maximum-bid related questions, the additional option was *To prevent selling*.
**entities to be ranked included: *banks*, *government*, *insurance companies*, *telcos*, *yourself*.

4.4.4 Data Statistics

The data used throughout this paper was collected from October 28th and December 11th 2013, inclusive.

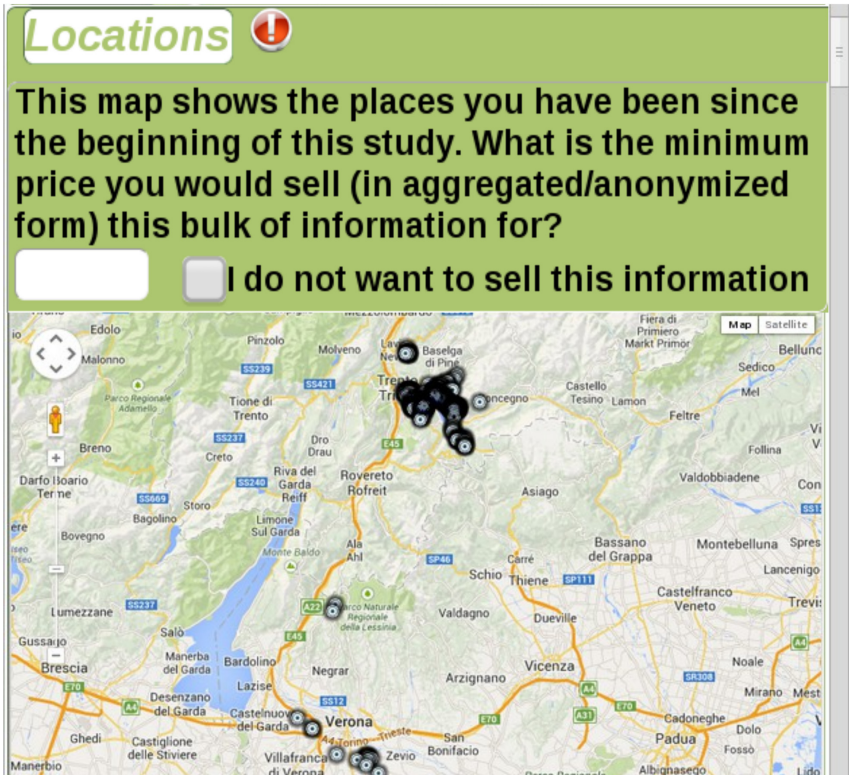


Figure 4.4: Location-specific bulk information question in the EoS survey.

Data was not collected for the first 3 days of November, due to the All Saints festivities in Italy; hence, our data-set encompasses 43 days. A total of 2838 daily surveys were administered during this period. Statistics on participation and bidding data follow.

Participation

The participation rates for daily surveys is 79%. As mentioned earlier, users were granted opt-out options for each survey question by ticking a check-box which portrayed "I do not want to sell this information". Table 4.21 reports statistics of opt-out and distributions of valid responses (*i.e.* survey items for which participants did not opt-out and entered their bid) for each category.

Category	Q_1	median	Q_3	mean	st_dev	OO (avg.%)	OO (median%)
Location	34.3	69.8	84.3	58.2	33.1	17.7	2.63
Communications	55.8	74.4	88.4	64.8	29.9	5.01	0
Running Apps	40.7	65.1	81.4	58	30.8	7.59	0
Media	62.8	76.7	90.7	66.4	32.8	9.25	0

Table 4.21: Distribution statistics of valid bid responses per category. Values reported in percentages. Last two columns portray the opt-out statistics per category.

Bids

Table 4.22 summarizes the bidding values for each personal data category and level of complexity. Values are included when the participant chose to assign an actual value, rather than opting out of the question. Figure 4.5 depicts median bid values each day for each category and level of complexity[6].

	Individual	Aggregated	Processed	Global
Location	[1, 3, 9]	[1, 3, 10]	[1, 2, 7]	[1, 3, 8]
Communications	[.95, 2, 5.96]	[1, 2, 8]	[.9, 2, 8]	[1, 2, 7]
Running Apps	[1, 2, 6]	[1, 1, 5]	[1, 2, 5]	[1, 2, 5]
Media	[.5, 1, 5]	[.5, 1, 5]	[.5, 1, 3]	[.5, 1, 4]

Table 4.22: [Q_1,median,Q_3] triplets for bid values (€) per category and level of complexity.

Awards

The total amount won by participants in the form of auction awards was €262 which was paid in Amazon vouchers. Additionally, we selected the ten subjects with the highest response rate and ran a raffle to select the winner of a final prize of €100.

A total of 29 subjects won at least one auction during the study; the cardinality of the winning set ramped from 5 to 29 as an effect of the increased number

[6]Note how the spatial gap between the first two interventions is smaller than between the rest of interventions because of the lack of data during 3 days in November.

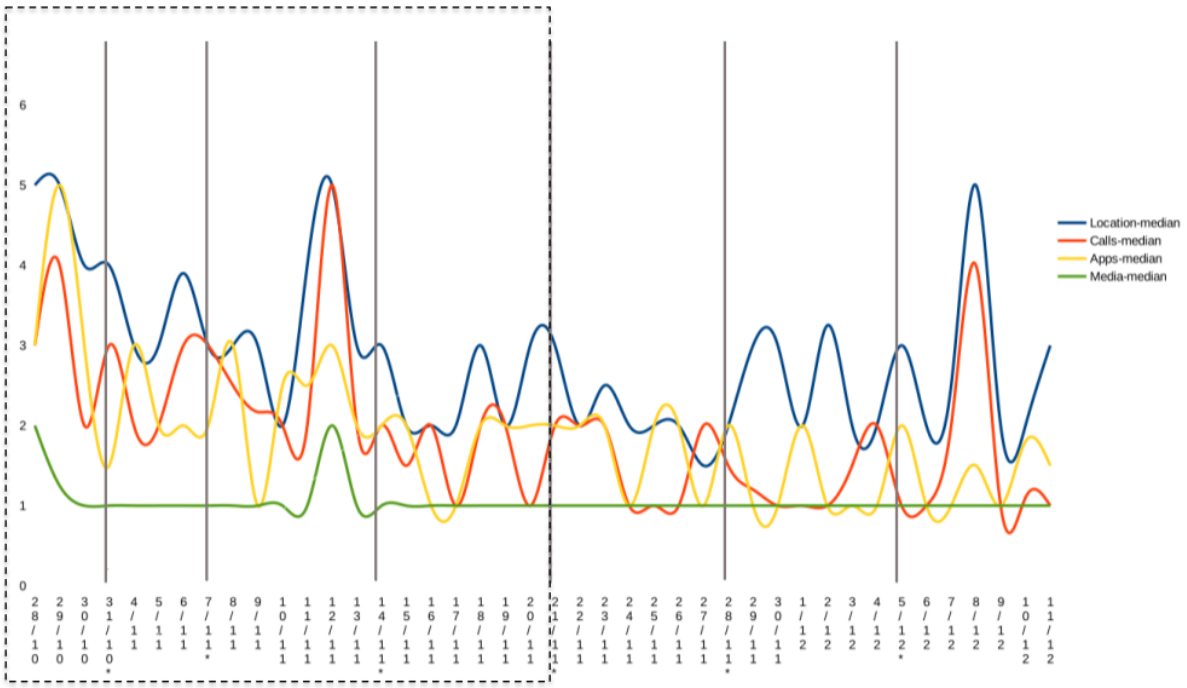


Figure 4.5: Daily median bid values per category. Vertical lines indicate interventions. Dashed area indicates phase 1.

of auctions run in the second phase of the study.

4.4.5 Data Analysis

The bidding data that was collected in the study is not normally distributed. Hence, we applied non-parametric analysis to test whether significant differences exist in the value distributions of different types of personal data. Thus, we report results using the Kruskal-Wallis test with a level of significance of $p < .05$.

Furthermore, we carried out correlation analyses to investigate whether associations between mobile phone usage patterns, demographics, subjects' predispositions, traits and auction behavior exist. For these analyses we employed the non-parametric Spearman's Rho method with a level of significance of $p < .05$.

Bids

We investigate first daily bids and specifically whether significant differences exist between (1) the categories and (2) levels of complexity within each category of mobile personal data we collected.

Between-Category Study Significant differences in bid distributions were found between all data categories, with the only exception of Communications and Apps.

The lack of statistically significant differences between Communications and Apps could be partially explained by the fact that most of the apps installed and used by participants in the study are communication apps. In terms of both running time and installations, $\sim$50% of the top 20 apps are messaging apps (WhatsApp and similar), email (Gmail, Hotmail, Y!Mail), voice-over-IP clients (Skype, Viber) and social networking clients (Facebook).

We thus hypothesize that the distinction between Communication and Apps might be blurred.

We leave the validation of such a hypothesis to future work. Nonetheless, the finding that participants seem to perceive, and consequently valuate, communications provided by a telco company and those provided by mobile apps in a similar manner, is intriguing and worth investigating.

Within-Category Study Next we analyzed the differences in the distribution of bids within the different levels of complexity of mobile personal data. In other words, we looked if bid distributions within a given mobile data category showed significant differences for individual, aggregated, and processed information.

Applications. Significant differences emerged between individual and aggregated information ($p = .0108$), and between aggregated and processed information ($p = .039$). In particular, aggregated information about running applications (e.g. *yesterday 7 applications were running on your device*) was valued

less ($\tilde{x} = €1$) than individual (e.g. *yesterday the Gmail application was running on your device*) or processed (e.g. *yesterday the Gmail application ran for 120 minutes on your device*) information ($\tilde{x} = €2$). No significant difference was found between monetary valuations of individual and processed information on running applications ($p = 0.659$).

Media. Within the Media category, a significant difference in bid distributions was found ($p = .046$) between aggregated (e.g. *yesterday you shot 8 pictures*) and processed (e.g. *yesterday night you shot 3 pictures*) information. While for both information types the median bid value is $\tilde{x} = €1$, a significant difference exists in terms of dispersion: the quartile coefficient of dispersion (*i.e.* the ratio between difference and sum of the 3rd and 1st quartiles) is, respectively, $qcod_{agg} = .81$ and $qcod_{pro} = .71$.

Communications and Locations. No significant differences were found in within-category analyses for Communications and Locations. In other words, participants valued similarly the communication and location data with each of the 3 levels of complexity.

Impact of the Change in Auction Strategy

As described earlier, in the middle of the study we increased the frequency of auctions from weekly (phase 1) to daily (phase 2). This change was transparent to participants and the frequency of email interventions was kept constant – every Thursday. We designed these two phases to assess if the probabilty of winning had an effect on bidding behavior.

Indeed, we observe significant differences in bid distributions between the two phases for all categories: locations (p=.02), communications (p=01), apps (p=.001) and media (p=.005). Moreover, we find that mobile PII valuations drop for all categories in the second phase, as more participants won the auctions to monetize their data.

The Value of Bulk Information

The monetary valuations gathered in the final questionnaire for bulk information (*i.e.* all the data collected in the 6-weeks presented in aggregated/anonymized form) are summarized in Table 4.23. Since participants could opt-out, we also report opt-out percentages for bulk information.

Comparing with daily bids (see Table 4.22), the median bids for bulk information are one order of magnitude larger than the median individual bids, except for the media category. Mean opt-out percentages are similar except for the apps category.

The value ranking obtained from daily bids (Location $>$ Communications $>$ Apps $>$ Media) is different from that obtained in bulk bids (Location $>$ Apps $>$ Communications $>$ Media). In particular, application-related bulk data is valued significantly higher than communications-related bulk information.

	Location	**Communications**	**Running Apps**	**Media**
mean	588.1	51.1	170.4	25.1
median	22.5	15	20	5
opt-out (%)	16.67	3.34	0	8.34

Table 4.23: Median/mean values (€) for bulk bids, and corresponding opt-out percentages.

Relationship between Bids and Daily Behaviors

In order to assess whether significant effects exist between mobile phone usage patterns and bidding behavior, we first computed daily behavioral variables from the sensed data. Table 4.24 depicts the variables that we extracted with a daily granularity and for each participant. With respect to *location* data, information about the number of places visited was derived under the assumption that two locations would correspond to different places if the distance between them was larger than a threshold set to 100 meters. The radius of gyration corresponds to the radius of the smallest circle encompassing all location points

registered each day.

For all these behavioral variables, we computed higher-order features corresponding to their statistical behavior over the 6-weeks period: mean, median, standard deviation, coefficient of variation (ratio of the standard deviation to the mean) and the quartile coefficient of dispersion. The last two features capture dispersion effects.

Furthermore, for each participant and data category, we computed mean, median, and standard deviation of their bids.

Category	**Daily Behavioral Variables**
Location	Distance *total/mean/median/std*
	Speed *mean/median/std*
	Radius of Gyration
	Number of Places Visited
Communications	Calls Duration *total/mean/median/std*
	Calls Diversity
	Calls Total
Applications	Total Apps Running
	Total Apps Running Time
Media	Total Pics shot

Table 4.24: Daily behavioral variables computed from mobile phone usage data.

Daily Bids We studied all correlations found between daily behavioral variables and bids in each category.

We found a positive correlation between the mean location bid value and the median of daily distance traveled ($R = .294, p = .024$). That is, the larger the daily distance traveled, the higher the valuations of location information. With respect to applications, there are several statistically significant correlations. In particular, the total app running time is negatively correlated with the median app bid value ($R = -.26, p = .048$), meaning that the more time a participant spent using mobile apps, the lower the median valuations of app information.

No significant correlation was found between communication and photo-taking behavioral features and bids on the communications and media categories.

Bulk Bids There were a number of significant correlations between bids on bulk information and daily behaviors. Below we summarize the most notable correlations that we found.

Mobility information was positively correlated with bids on bulk *location*, *communication* and *application* information. In particular, with the median of the i) radius of gyration ($R = .46, p = .0008$ for loc.; $R = .37, p = .005$ for comm.; $R = .34, p = .009$ for apps); and ii) daily mean speed ($R = .29, p = .04$ for loc.; $R = .39, p = .002$ for comm.; $R = .29, p = .029$ for apps). Location and application data was also positively correlated with the median of the daily mean distance traveled ($R = .39, p = .005$ for loc.; $R = .28, p = .031$ for apps) whereas communication bids were also positively correlated with the median of the i) total distance traveled ($R = .314, p = .018$) and ii) number of places visited ($R = .336, p = .011$).

We also found statistically significant negative correlations of bulk location, communication and application bids with the coefficients of variation of mobility variables.

These correlations imply that the larger the daily distance traveled, the higher the valuation of location, communication and application bulk bids. Conversely, the higher the variation in the patterns of mobility of a person, the lower his/her valuation of location, communication and app bulk information.

Note that bulk communication bids were not correlated with communication variables. In addition, bulk application bids are negatively correlated with the cumulative sum of daily unique total apps ($R = -.37, p = .003$) and with the median ($R = -.28, p = .029$) and mean ($R = -.26, p = .04$) of total apps running daily.

Finally, bulk media bids are correlated with the cumulative sum of daily

unique total apps ($R = -.29, p = .03$).

Relationship between Bids, Demographics, Traits and Dispositions

Daily Bids In the case of daily bids, we did not find any meaningful statistically significant correlation between bids and our participants' demographics or personality.

There were statistically significant correlations with *self-disclosure* variables that could be explained by the relevance of privacy aspects for all types of self-disclosure [192]. In particular, the Intentional/Unintentional factor in self-disclosure is positively correlated with bids in three categories (communication, applications and media): (1) mean ($R = .258, p = .048$), median ($R = .291, p = .02$) and standard deviation ($R = .323, p = .012$) in communication bids, (2) median application bid value ($R = .26, p = .04$), and (3) median ($R = .30, p = .02$), mean ($R = .27, p = .041$), and standard deviation ($R = .305, p = .019$) of media bids.

Bulk Bids Bulk location bids are found to be negatively correlated with Creativity ($R = -.375, p = .007$), while having positive correlations with the Intentional/Unintentional factor in self-disclosure ($R = .295, p = .039$) and Agreeableness ($R = .31, p = .027$). Interestingly, a positive correlation exists between bulk location bids and personal income ($R = .32, p = .02$). Furthermore, bulk communication information positively correlates with Agreeableness ($R = .31, p = .018$), and with the Intentional/Unintentional factor in self-disclosure ($R = .34, p = .009$).

4.4.6 Insights from the EoS survey

In the final survey, we asked our participants about particular bids they made during the 6-week data collection phase, and gave them the opportunity to ex-

press their views and concerns in free-form text (see Table 4.20 for details of the EoS survey questions).

Trust

As seen in Table 4.20, Q4 asked our participants about their trust preferences with respect to 5 different entities who could be the safekeepers of their personal data: themselves, banks, telcos, governments and insurance companies. From the trust rankings provided by our participants, we computed a *trust score* for each entity by assigning a 1 to 5 value according to its rank and subsequently normalizing by the number of respondents. The final ranking that we obtained was: *yourself* (.997), *banks* (.537), *telcos* (.513), *government* (.49), and *insurance companies* (.46). This result is aligned with the initial survey answers (Q2 and Q3 in Table 4.19) where participants conveyed that they do not trust telco operators or app providers with how they use their data.

In sum, overwhelmingly our participants trust themselves with their personal data more than any other entity, followed by banks and telcos. Insurance companies were the least trusted party. A similar question was also asked by Carrascal *et al.* [39] obtaining similar results: the most trusted entity for a subject was the subject himself and the least trusted entities were the insurance companies. Interestingly, in our study, conducted in Italy, government was the second *least* trusted entity while in Carrascal *et al.* [39], conducted in Spain, the government was the second *most* trusted entity.

Lowest/Highest Bids per Category

When analyzing the lowest/highest bids per category, we found that 70% of the highest bids for all categories took place in the first phase of the study (during the first three weeks). Adding more auctions (as it happened in the second phase of the study) led to lower bids.

In the communications category, 61% of the time participants entered a low bid to win and sell the associated communications information. This was significantly higher than for any other category. For all other categories, the most common reason reported for entering the low bid was that the information was not important. This finding suggests that participants found communication data to be the most desirable to sell.

Conversely, location was the most sensitive category of information as 25% of the time participants entered a high location bid in order to avoid selling the information. This was significantly higher than for the other categories (5% for communications, 3% for apps and 6% for media).

Insights about Opt-out Choices

Location was the category of data for which subjects opted-out the most (56%), followed by media (24%), apps (18%) and communications (2%). In the free-text explanations provided by our subjects it is clear that location is deemed to be the most sensitive category of information, *e.g.*:

> *I don't like the idea of being geo-localized.*
> *This kind of information is too detailed and too personal.*

Interesting explanations were also provided to justify the choice of not selling apps information, including that from apps usage is possible to infer information related to interests, opinions (expecially political opinions), and tastes:

> *From the usage of some applications it is possible infer information such as political orientation and other opinions and interests.*

4.4.7 Discussion and Implications

From the previously described analyses we can draw six insights related to mobile personal data:

1. The Value of Bulk Mobile PII: Carrascal *et al.* [39] have reported higher values in their study on valuation of personal Web-browsing information than the ones we obtained in our study. The overall median bid value in our study was $\tilde{x} = €2$ while Carrascal *et al.* reported an overall median bid value equal to $\tilde{x} = €7$ when they took in account context-dependent personal information. There are a few methodological differences between both studies which might explain the differences in bid values. In particular, [39] asked participants to provide a valuation of personal information captured while browsing the Web *in-situ* using a rSEM methodology. Instead, we employed a DRM methodology querying participants about ther mobile PII from the previous day.

From the valuations obtained in [39] and our study, it seems that individual pieces of PII are not as valuable when queried *out-of-context* –such as in our study– than *in context* –such as in [39].

Conversely, bulk mobile PII was valued higher in our study than in [39] and significantly higher than individual PII. As shown in Tables 4.22 and 4.23, bulk information was valued an order of magnitude higher than individual data except for information in the media category. This finding is probably due to the power of the visualizations in the EoS survey, particularly for location and apps data. One hypothesis for this higher valuation is that participants realized how bulk data conveyed information about their life-style and habits and therefore considered it to be more valuable than daily items. Recently, Tang *et al.* have shown the impact of different visualization types (text-, map-, and time-based) on social sharing of location data [273].

This result has a direct consequence for the design of trading mobile PII and highlights an asymmetry between buyers and sellers: for buyers, it would be more profitable to implement mechanisms to trade single pieces of information –that they could later aggregate. For sellers, however, it would be more advantageous to sell bulks of information.

2. Location, location, location: As shown in Tables 4.22 and 4.23, location

information received the highest valuation for all levels of complexity and was the most opted-out category of mobile PII. Bulk location information was very highly valued, probably due to the powerful effect of the map visualization in the EoS survey. Several participants also expressed that they did not want to be geolocalized and considered location information to be highly sensitive and personal.

Moreover, we found statistically significant correlations between mobility behaviors (*e.g.* mean daily distance traveled, daily radius of gyration, etc.) and valuations of personal data. Not all users value their personal data equally: the more someone travels on a daily basis, the more s/he values not only her/his location information but also her/his communication and application information. These insights may have an impact on the design of commercial location-sharing applications. While users of such applications might consent at install time to share their location with the app, our work suggests that when explicitly asked about either individual or bulk location data, $\sim 17\%$ of users decide not to share their location information. In addition, mobility behaviors will influence the valuations of PII.

Tsai *et al.* [280] conducted an online survey with more than 500 American subjects to evaluate their own perceptions of the likelihood of several location-sharing scenarios along with the magnitude of the benefit or harm of each scenario (*e.g.* being stalked or finding people in an emergency). The majority of the participants found the risks of using location-sharing technologies to be higher than the benefits.

However, today a significant number of very popular mobile apps such as Foursquare and Facebook Places make use of location data. These popular commercial location sharing apps seem to mitigate users' privacy concerns by allowing them to selectively report their location using check-in funtionalities instead of tracking them automatically.

Based on our findings and given our participants concerns and high valua-

tions of bulk location information, we believe that further user-centric studies on sharing and monetary valuation of location data are needed.

3. Socio-demographic characteristics do not matter, behavior does: When we correlated bid values against socio-demographic characteristics, we did not find significant correlations. This result is in contrast to previous work that found socio-demographic (mainly sex and age) differences in privacy concerns and consequent information disclosure behaviors [86, 48].

However, these previous studies were focused mainly on online information and on disclosure attitudes and privacy concerns than on monetary valuation of personal data. Carrascal *et al.* [39], instead, found results in line with ours (no significant correlations) except for a surprising low valuation of online information from older users.

Conversely, we found statistically significant correlations between behavior (particularly mobility and app usage) and valuations of bids. From our findings it seems that personal differences in valuations of mobile PII are associated with behavioral differences rather than demographic differences. In particular, the larger the daily distance traveled and radius of gyration, the higher the valuation of PII. Conversely, the more apps a person used, the lower the valuation of PII. A potential reason for this correlation is due to the fact that savvy app users have accepted that mobile apps collect their mobile PII in order to provide their service and hence value their mobile PII less.

4. Intentional self-disclosure leads to higher bids: We found a positive correlation between the Intentional/Unintentional dimension of self-disclosure and the median values of the bids. This result could be explained by the fact that people with more intentional control about disclosing their own personal information, may be more aware of their personal data and hence also value it more from a monetary point of view.

Interestingly, we did not find significant correlations between bid values and other traits with the exception of Agreeablenness and bulk location and com-

munication bids. Previous studies on the influence played by individual traits (usually personality traits and LoC) on privacy dispositions and privacy-related behaviors have provided contrasting evidence: some of them found small correlations [176, 229], while Schrammel *et al.* found no correlations [245]. Hence, our results require additional investigations in order to clarify which are, if any, the dispositions and individual characteristics to take in account when a buyer does a monetary offer for personal data.

5. Trust: From our study and from Carrascal *et al.* [39], it clearly emerges that invividuals mainly trust themselves to handle their own personal data. This result suggests the adoption of a decentralized and *user-centric* architecture for personal data management. Recently, several research groups have started to design and build personal data repositories with functionalities that enable people to control, collect, delete, share, and sell personal data [66, 199] whose value to users is supported by our findings.

6. Unusual days lead to higher bids: During our study there were two unusual days: December 8th (Immaculate Conception Holiday) and November 11th (a day with extremely strong winds which caused multiple road blocks and accidents). As can be seen in Figure 4.5, the median bids for all categories in these two days were significantly higher than for the rest of the days in the study. Perhaps not surprisingly, participants in our study value their PII higher in days that are unsual when compared to typical days.

This result suggests that not all PII even within the same category and level of complexity is valued equally by our participants, which has a direct implication for personal data markets and for services that monetize mobile personal data.

4.4.8 Conclusion

We have investigated the monetary value that people assign to their PII as it is collected by their mobile phone. In particular, we have taken into account four categories of PII (location, communication, apps and media) with three

levels of complexity (individual, aggregated and processed). We have carried out a comprehensive 6-week long study in a living lab environment with 60 participants adopting a Day Reconstruction Method along with a reverse second price auction mechanism to collect honest monetary valuations.

We have found that location is the most valued category of PII and that bulk information is valued much higher than individual information (except for the media category). We have identified individual differences in bidding behaviors which are not correlated with socio-demographic traits, but are correlated with behavior (mobility and app usage) and intentional self-disclosure. Finally, we have found that participants trust themselves with their PII above banks, telcos and insurance companies and that unusual days are perceived as *more valuable* than typical days.

Chapter 5

Harvesting the Wild Wild (and Social) Web

Reactions to posts in an online social network show different dynamics depending on several textual features of the corresponding content. Do similar dynamics exist when images are posted? Exploiting a novel dataset of posts, gathered from the most popular Google+ users, we try to give an answer to such a question in Section 5.1. We describe several virality phenomena that emerge when taking into account visual characteristics of images (such as orientation, mean saturation, etc.). We also provide hypotheses and potential explanations for the dynamics behind them, and include cases for which common-sense expectations do not hold true in our experiments.

Furthermore, we noticed how while many lexica annotated with words polarity are available for sentiment analysis, very few tackle the harder task of emotion analysis and are usually quite limited in coverage. In Section 5.2, we present a novel approach for extracting – in a totally automated way – a high-coverage and high-precision lexicon of roughly 37 thousand terms annotated with emotion scores, called `DepecheMood`. Our approach exploits in an original way 'crowd-sourced' affective annotation implicitly provided by readers of news articles from `rappler.com`. By providing new state-of-the-art performances in unsupervised settings for regression and classification tasks, even

using a naïve approach, our experiments show the beneficial impact of harvesting social media data for affective lexicon building and emotion analysis.

5.1 Exploring Image Virality in Google Plus

How do things become 'viral' on the Internet? And what exactly do we mean by 'influence'? Since marketing and industry people want their messages to spread in the most effective and efficient way possible, these questions have received a great deal of attention, particularly in recent years, as we have seen a dramatic growth of social networking on the Web. Generally speaking, virality refers to the tendency of a content either to spread quickly within a community or to receive a great deal of attention by it. In studying the spreading process we will focus on the content and its characteristics, rather than on the structure of the network through which the information is moving. In particular, we will investigate the relationships between visual characteristics – of images enclosed in Google+ posts – and virality phenomena. We will use three virality metrics: plusoners, replies and resharers.

This exploratory work stems from the use people make of social networking websites such as Google+, Facebook and similar: we hypothesized that perceptual characteristics of an image could indeed affect the virality of the post embedding it, and that – for example – cartoons, panorama or self-portraits picture affect users' reactions in different ways. The aim of this work is to investigate whether signs of such "common-sense" intuition emerge from large-scale data made available on popular social networking websites like Google+ and, in such case, to open discussion on the associated phenomena.

5.1.1 Related Works

Several researchers studied information flow, community building and similar processes using Social Networking sites as a reference [1, 132, 148, 169]. How-

ever, the great majority concentrates on network-related features without taking into account the actual content spreading within the network [168]. A hybrid approach focusing on both product characteristics and network related features is presented in [13]: the authors study the effect of passive-broadcast and active-personalized notifications embedded in an application to foster word of mouth.

Recently, the correlation between content characteristics and virality has begun to be investigated, especially with regard to textual content; in [131], for example, features derived from sentiment analysis of comments are used to predict the popularity of stories. The work presented in [27] uses *New York Times*' articles to examine the relationship between emotions evoked by the content and virality, using semi-automated sentiment analysis to quantify the affectivity and emotionality of each article. Results suggest a strong relationship between affect and virality; still, the virality metric considered is interesting but very limited: it only consists of how many people emailed the article. The relevant work in [60] measures a different form of content spreading by analyzing which are the features of a movie quote that make it "memorable" online. Another approach to content virality, somehow complementary to the previous one, is presented in [249], trying to understand which modification dynamics make a meme spread from one person to another (while movie quotes spread remaining exactly the same).

More recently, some works tried to investigate how different textual contents give rise to different reactions in the audience: the work presented in [103] correlates several viral phenomena with the wording of a post, while [101] show that specific content features variations (like the readability level of an abstract) differentiate among virality level of downloads, bookmarking, and citations. Still, to our knowledge, no attempt has been made yet to investigate the relation between visual content characteristics and virality.

5.1.2 Data Description

Using the Google+ API[1], we harvested the public posts from the 979 top followed users in Google+ (`plus.google.com`), as reported by the `socialstatistics.c` website on March 2nd 2012[2]. The time span for the harvesting is one year, from June 28th 2011 (Google+ date of launch) to June 29th 2012.

We decided to focus on the most popular users for several reasons: (i) the dataset is uniform from the point of view of sample role, i.e. VIPs, (ii) the behavior of the followers is consistent – e.g. no friendship dynamics – and (iii) extraneous effects due to followers network is minimized, since top followed users' network is vast enough to grant that, if a content is viral, a certain amount of reactions will be obtained.

We defined 3 subsets of our dataset, comprising respectively: (i) posts containing a static image, (ii) posts containing an animated image (usually, `gif`), (iii) posts without attachments (text-only). All other posts (containing as attachment videos, photo albums, links to external sources) were discarded. Statistics for our dataset are reported in Table 5.1. For each post, we considered three virality metrics[3]:

- **Plusoners**: the number of people who +1'd;
- **Replies**: the number of comments;
- **Resharers**: the number of people who reshared.

In Figures 5.1 and 5.2 we display the evolution over time of the network underlying our dataset (using a week as temporal unit), and of the reactions to posts given by users, respectively. We notice that:

[1] https://developers.google.com/+/api/

[2] The dataset presented and used in this work will be made available to the community for research purposes.

[3] Since the API provide only an aggregate number, we cannot make any temporal analysis of how reactions to a post were accumulated over time.

Table 5.1: An overview of the Google+ dataset.

Global	
actors	979
posts	289434
published interval	6/28/11–6/29/12
Posts with static images	
actors	950
posts	173860
min/max/median posts per actor	1/3685/ 65.5
min/max/median plusoners per post	0/9703/33.0
min/max/median replies per post[a]	0/571/12.0
min/max/median resharers per post	0/6564/4.0
Posts with animated images	
actors	344
posts:	12577
min/max/median posts per actor	1/2262/3.0
min/max/median plusoners per post	0/5145/17.0
min/max/median replies per post	0/500/7.0
min/max/median resharers per post	0/6778/10.0
Posts without attachments, text-only	
actors	939
posts	102997
min/max/median posts per actor	1/1744/41.0
min/max/median plusoners per post	0/20299 /16.0
min/max/median replies per post	0/538/17.0
min/max/median resharers per post	0/13566/1.0

[a] Replies count is cut around 500 by the API service.

1. the average number of reactions per user shows quite different trends depending on the metric considered: while replies tend not to be affected by the growth of the network, reshares and, to a lesser degree plusones, show an ever-growing trend.

2. The temporal plot of the average number of followers per user (Figure 5.1) in our dataset (in Google+ terminology, the number of people who *circled* them) shows a gradient increase around weeks 28/29. Interestingly, this is reflected in the plot of reactions over time (Figure 5.2): the gradient increases around the same weeks, for reshares and plusones; these effects are most probably due to Google+ transitioning from beta to public in late September 2011 (a similar phenomenon is reported also in [244]).

3. Finally, the orders of magnitude of such growths are very different: we notice that while reactions increase of a factor of 7 over the time period we took into account, the total number of followers increased of a factor of 25.

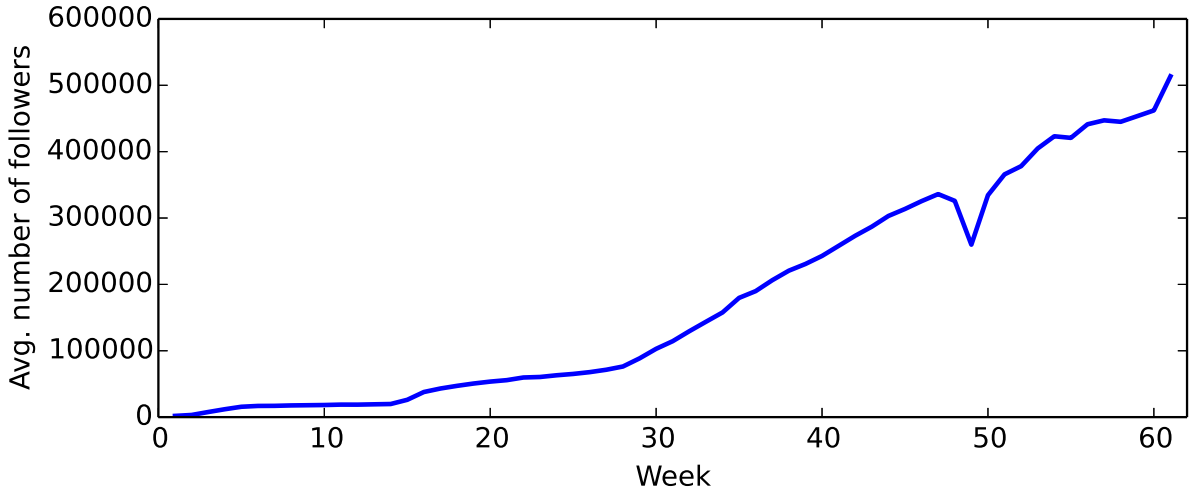


Figure 5.1: Average number of followers per user, at 1-week temporal granularity.

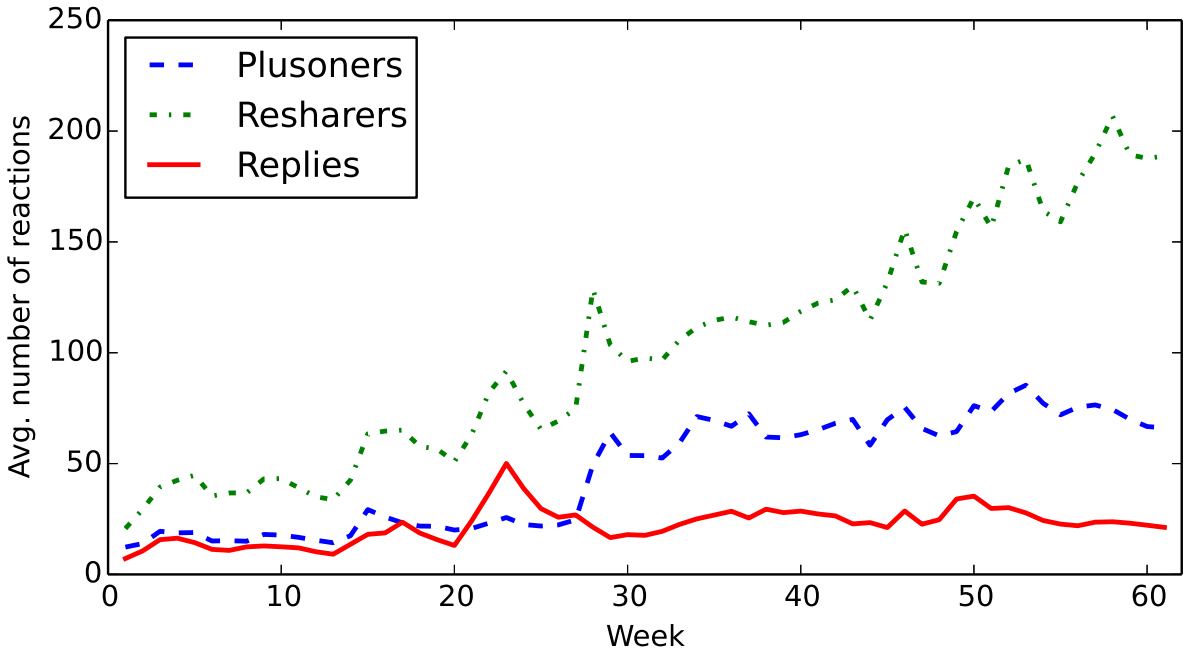


Figure 5.2: Average number of reactions per user, at 1-week temporal granularity. This value represents the average number of reactions elicited by each user's posts over 1-week time-slices.

The relative amount of followers' reactions does not significantly increase as the network grows[4]. As detailed in the next section, our analyses are based on comparing probability distributions: e.g. we evaluate if grayscale images have a significantly higher or smaller probability of reaching a certain virality score than colored ones. In the following analyses, for the sake of clarity, our discussion will not take into account the normalization factor (i.e. the size of the audience when a content is posted).

Indeed, we have run the same analyses normalizing the virality indexes of a given post against its *potential* audience: i) the effects are still visible, ii) the effects are consistent both in significance and sign with the not-normalized distributions, but iii) differences have lower magnitude (explained by the fact that virality indexes should be normalized using the *actual* audience – e.g. the

[4]It has been noted how (see, for instance, http://on.wsj.com/zjRr06), especially in the time frame we consider, users' activity did not increase much in front of the exploding network size.

followers exposed to the content).

Thus, since we are interested in comparing the virality of different image categories and our preliminary experiments showed that by normalizing the indexes their comparisons, their sign, and the derived interpretations still hold, we choose to report the non-normalized version of the results that are more intuitively readable.

In the following sections, after the analyses of text-only posts and of posts containing an animated image, we will consider the subset of static images as the reference dataset. Exemplar pictures taken from the dataset are shown in Figure 5.3, depicting some image categories that we will take into account in the following sections.

5.1.3 Data Analysis

Virality metrics in our dataset follow a power-law-like distribution thickening toward low virality score. In order to evaluate the "virality power" of the features taken into account, we compare the virality indexes in terms of empirical Complementary Cumulative Distribution Functions (CCDFs). These functions are commonly used to analyse online social networks in terms of growth in size and activity (see for example [7, 134], or the discussion presented in [170]) and also for measuring content diffusion, e.g. the number of retweets of a given content [156]. Basically, these functions account for the probability p that a virality index will be greater than n and are defined as follows:

$$\hat{F}(n) = \frac{\text{number of posts with virality index} > n}{\text{total number of posts}} \tag{5.1}$$

For example, the probability of having a post with more than 75 plusoners is indicated with $\hat{F}_{plus}(75) = \mathrm{P}(\#\text{plusoners} > 75)$. In the following sections we use CCDFs to understand the relation between image characteristics and post virality; in order to assess whether the CCDFs of the several types of posts we

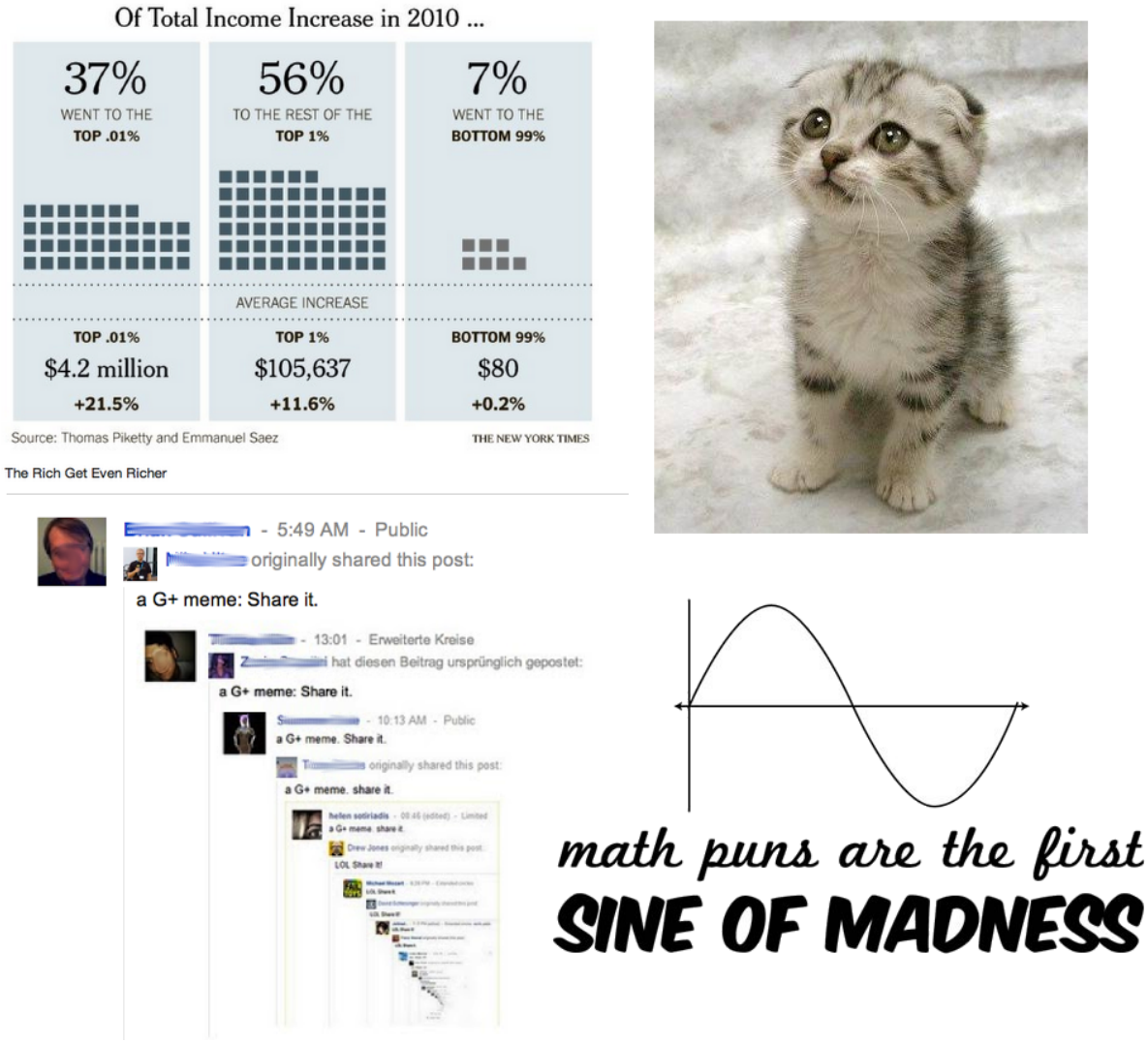


Figure 5.3: Exemplar pictures from the dataset.

take into account show significant differences, we will use the Kolmogorov-Smirnov (K–S) goodness-of-fit test, which specifically targets cumulative distribution functions.

Image vs. text-only

First of all, we aim to understand what is the impact of "adding an image to a post" in Google+. Some studies [279] already show that posts containing

an image are much more viral than simple plain-text posts, and that various characteristics of image based banners affect viewer's recall and clicks [172]. This finding can be explained in light of a "rapid cognition" model [9, 147]. In this model, the user has to decide in a limited amount of time, and within a vast information flow of posts, whether to take an action on a particular post (e.g. to reply, reshare, give it a plusone). Thus, pictures, and the characteristics thereof analyzed in the following sections, might play a role of paramount importance in her decision-making process as she exploits visual cues that grab her attention. In some respects, the rapid cognition model is reminiscent of the mechanisms by which humans routinely make judgments about strangers' personality and behavior from very short behavioral sequences and non-verbal cues [58, 167].

In order to investigate the general impact of images we compared posts containing a picture with posts containing only text. While our findings overall coincide with [279], some interesting phenomena emerged. First, we see that the probability for a post with an image to have a high number of resharers is almost three times greater ($\hat{F}_{resh}(10)$ = 0.28 vs. 0.10, K–S test $p < 0.001$), see Figure 5.4.c. Still, the CCDFs for the other virality indexes show different trends:

- Posts containing images have lower probability of being viral when it comes to number of comments ($\hat{F}_{repl}(50)$ = 0.33 vs. 0.22, K–S test $p < 0.001$), see Figure 5.4.b. This can be explained by the fact that text-only posts elicit more "linguistic-elaboration" than images (we also expect that the average length of comments is higher for text-only posts but we do not investigate this issue here).

- Also, if we focus on simple appreciation (plusoners in Figure 5.4.a), results are very intriguing: while up to about 75 plusoners the probability of having posts containing images is higher, after this threshold the situa-

tion capsizes. This finding can be of support to the hypothesis that, while it is easier to impress with images in the information flow — as argued with the aforementioned "rapid cognition" model — high quality textual content can impress more.

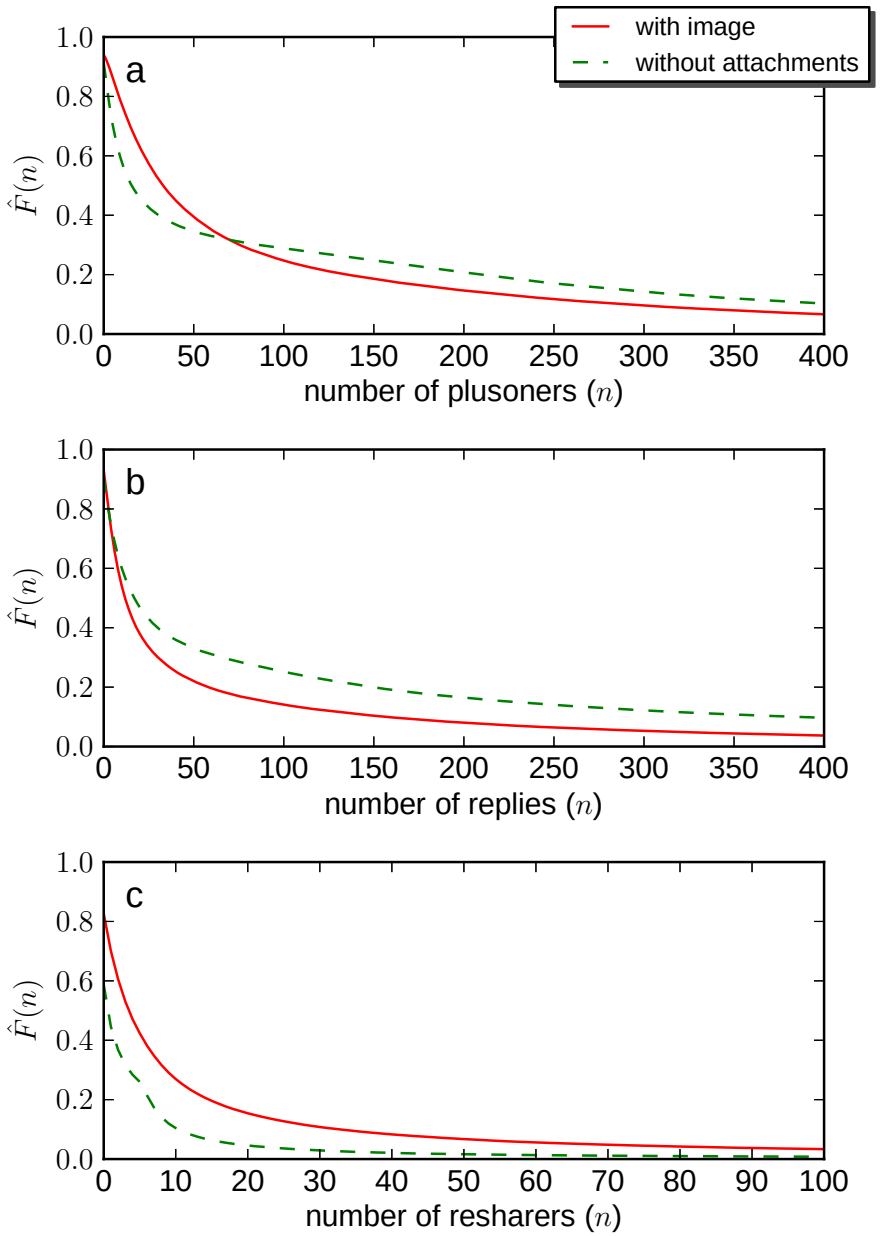


Figure 5.4: Virality CCDFs for posts with image vs. text-only posts.

Static vs. Animated

Animated images add a further dimension to pictures expressivity. Having been around since the beginning of the Internet (the `gif` format was introduced in late 80's), animated images have had alternate fortune, especially after the wide spread of services like youtube and the availability of broadband. Nonetheless, they are still extensively used to produce simple animations and short clips. Noticeably, the value of simple and short animations has been acknowledged by Twitter with the recently released *Vine* service.

Whether a post contains a static or animated image has a strong discriminative impact on all virality indexes, see Figure 5.5. With respect to plusoners and replies, static images tend to show higher CCDFs (respectively two and three times more, $\hat{F}_{plus}(75)$ = 0.30 vs. 0.17, $\hat{F}_{repl}(50)$ = 0.22 vs. 0.08, K–S test $p < 0.001$), while on resharers the opposite holds.

The fact that $\hat{F}_{resh}(n)$ is two times higher for posts containing animated images ($\hat{F}_{resh}(10)$ = 0.48 vs. 0.27, K–S test $p < .001$) can be potentially explained by the fact that animated images are usually built to convey a small "memetic" clip - i.e. *funny*, *cute* or *quirky* situations as suggested in [74].

In order to verify this hypothesis we have annotated a small random subsample of 200 images. 81% of these animated images were found to be "memetic" (two annotators were used, positive example if the image score 1 at least on one of the aforementioned dimensions, annotator agreement is very high — Cohen's kappa 0.78). These findings indicate that animated images are mainly a vehicle for amusement, at least on Google+.

Image Orientation

We then focused on the question whether image orientation (*landscape*, *portrait* and *squared*) has any impact on virality indexes. We included squared images in our analysis since they are typical of popular services a la *Instagram*. These

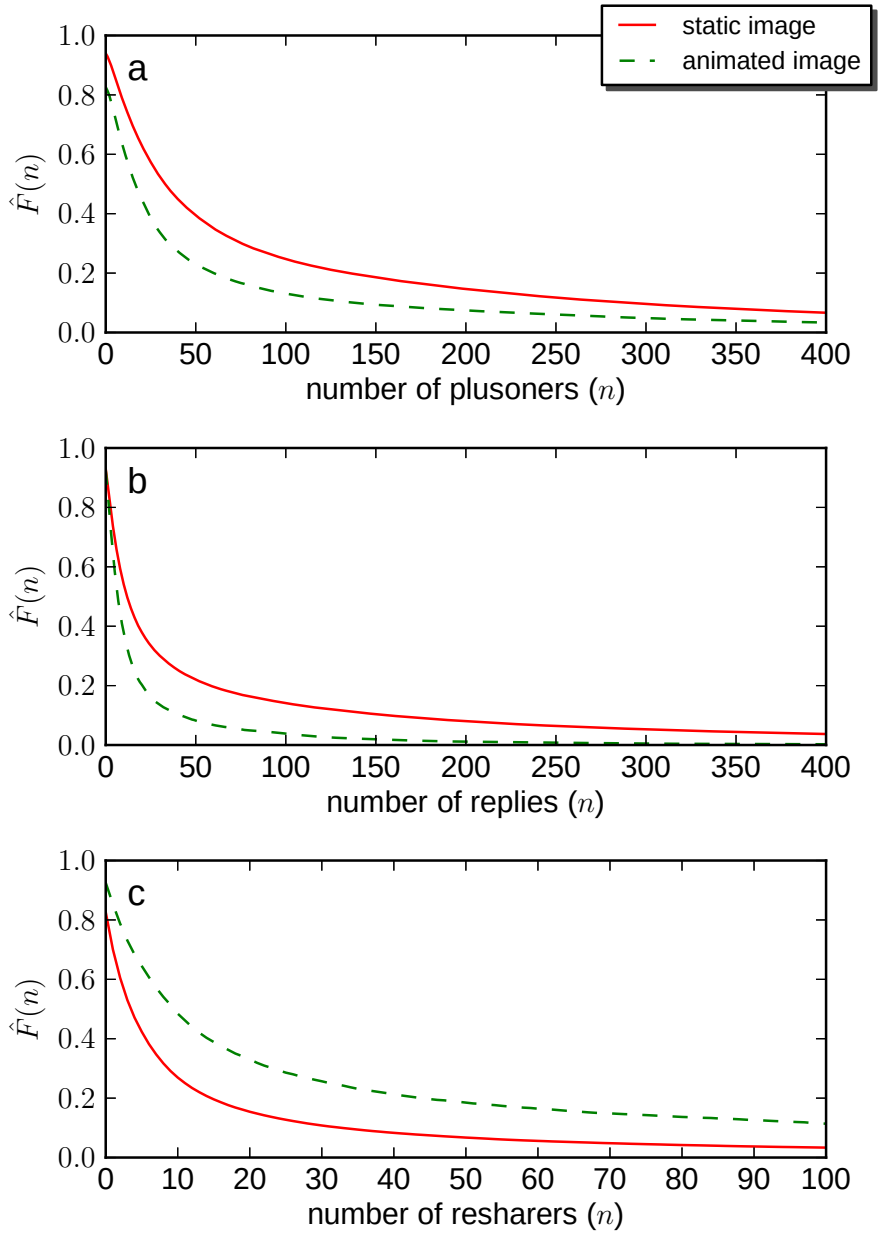


Figure 5.5: Virality CCDFs for static vs. animated images.

services enable users to apply digital filters to the pictures they take and confine photos to a squared shape, similar to Kodak Instamatic and Polaroids, providing a so-called "vintage effect".

We have annotated a small random subsample of 200 images. 55% of these

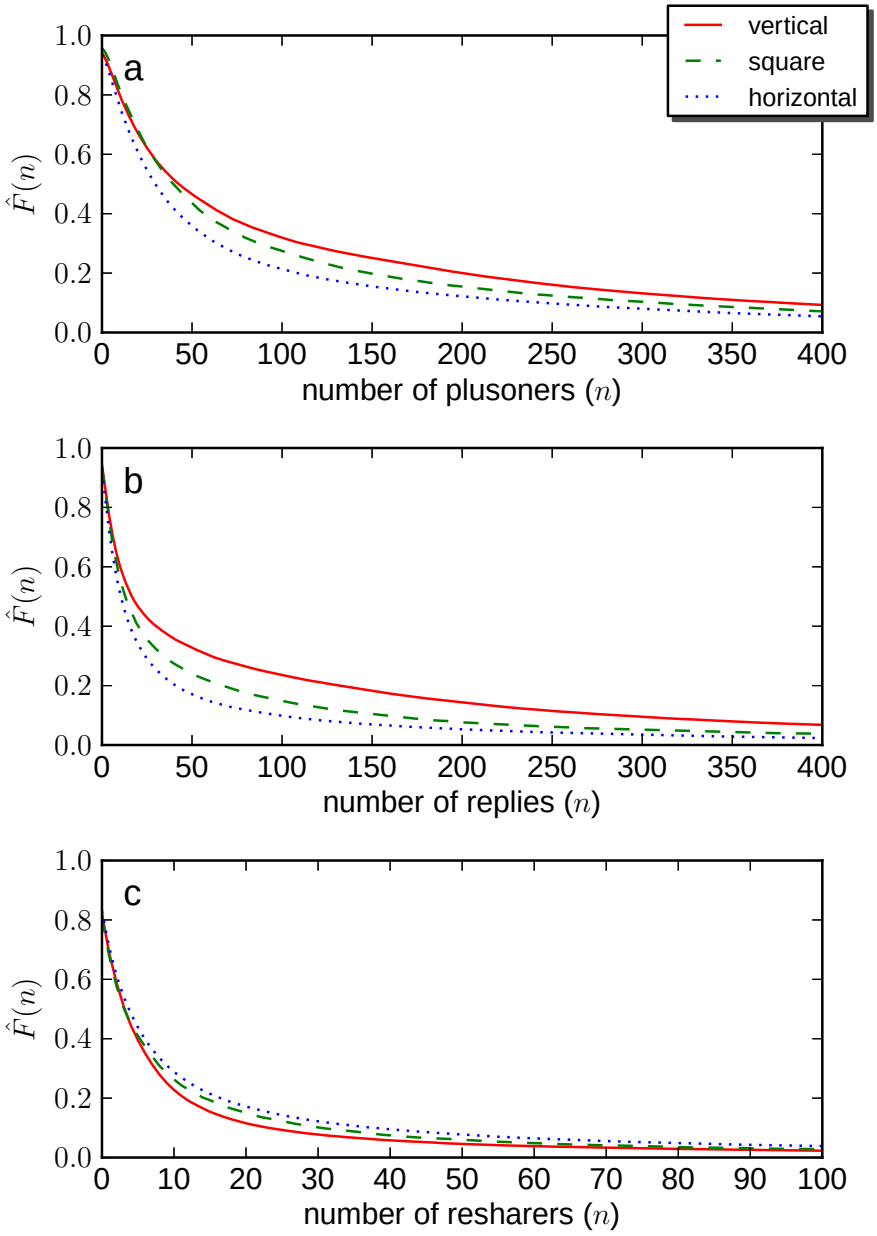


Figure 5.6: Virality CCDFs for image orientation.

images were found to be "Instagrammed" (two annotators were used, positive example if the image is clearly recognized as modified with a filter; annotator agreement is high – Cohen's kappa 0.68). Note that, if we include also black and white squared pictures without any other particular filter applied (b/w is one of

the "basic" filter provided by *Instagram*) the amount of Instagrammed pictures rises to 65%. Obviously, the ratio of pictures modified with this and similar services could be higher; here, we rather wanted to identify those pictures that were clearly recognized as seeking for the aforementioned "vintage effect".

While the orientation seems not to have strong impact on resharers, with a mild prevalence of horizontal pictures (see Figure 5.6.c), plusoners and replies tend to well discriminate among various image orientations. In particular, portrait images show higher probability of being viral than squared images than, in turn, landscapes (see Figure 5.6.a and 5.6.b).

Furthermore, CCDFs indicate that vertical images tend to be more viral than horizontal ones ($\hat{F}_{plus}(75)$ = 0.38 vs. 0.26, $\hat{F}_{repl}(50)$ = 0.38 vs. 0.17, K–S test $p < 0.001$). Hence, while squared images place themselves in the middle in any metric, landscape images have lower viral probability for plusones and replies but slightly higher probability for reshares.

This can be partially explained by the fact that we are analyzing "celebrities" posts. If the vertically-orientated image contains the portrait of a celebrity this is more likely to be appreciated rather that reshared, since the act of resharing can also be seen as a form of "self-representation" of the follower (we will analyze the impact of pictures containing faces in the following section). The opposite holds for landscapes, i.e. they are more likely to be reshared and used for self-representation.

Images containing one face

In traditional mono-directional media (e.g. tv, billboards, etc.) a widely used promotion strategy is the use of testimonials, especially celebrities endorsing a product. Is the same strategy applicable to Social Media? Understanding the effect of posting images with faces by most popular Google+ users (and hypothesizing that those are their faces) is a first step in the direction of finding an answer.

We computed how many faces are found in the images, along with the ratio of the area that include faces and the whole image area, using the Viola-Jones [292] face detection algorithm. We considered images containing one face vs. images containing no faces. We did not consider the surface of image occupied by the face (i.e. if it is a close-up portrait, or just a small face within a bigger picture). The discriminative effect of containing a face on virality is statistically significant but small. Still, the pictures containing faces tend to have mild effect on resharers (slightly higher replies and plusoners but lower resharers as compared to images with no faces).

In order to verify the hypothesis mentioned earlier, i.e. that self-portraits tend to be reshared less, we also focused on a subsample of images containing faces that cover at least 10% of the image surface (about 6400 instances). In this case, the differences among indexes polarize a little more (higher plusoners and comments, lower resharers), as we were expecting. Unfortunately, images with even higher face/surface ratio are too few to further verify the hypotheses.

Grayscale vs. Colored

The impact and meaning of black-and-white (i.e. grayscale) photographic images has been studied from different perspectives (e.g. semiotics and psychology) and with reference to different fields (from documentary to arts and advertising). Rudolf Arnheim, for example, argues that color produces essentially *emotional* experience, whereas shape corresponds to *intellectual* pleasure [17]. Hence, black-and-white photography, because of its absence of expressive colors, focuses on shapes that require intellectual reflection and brings to explore aesthetic possibilities. We want to understand if such functions and effects can be spotted in our virality indexes.

In order to have a "perceptual" grayscale (some images may contain highly desaturated colors and so perceived as shades of gray) we dichotomized the dataset according to the mean-saturation index of the images, using a very con-

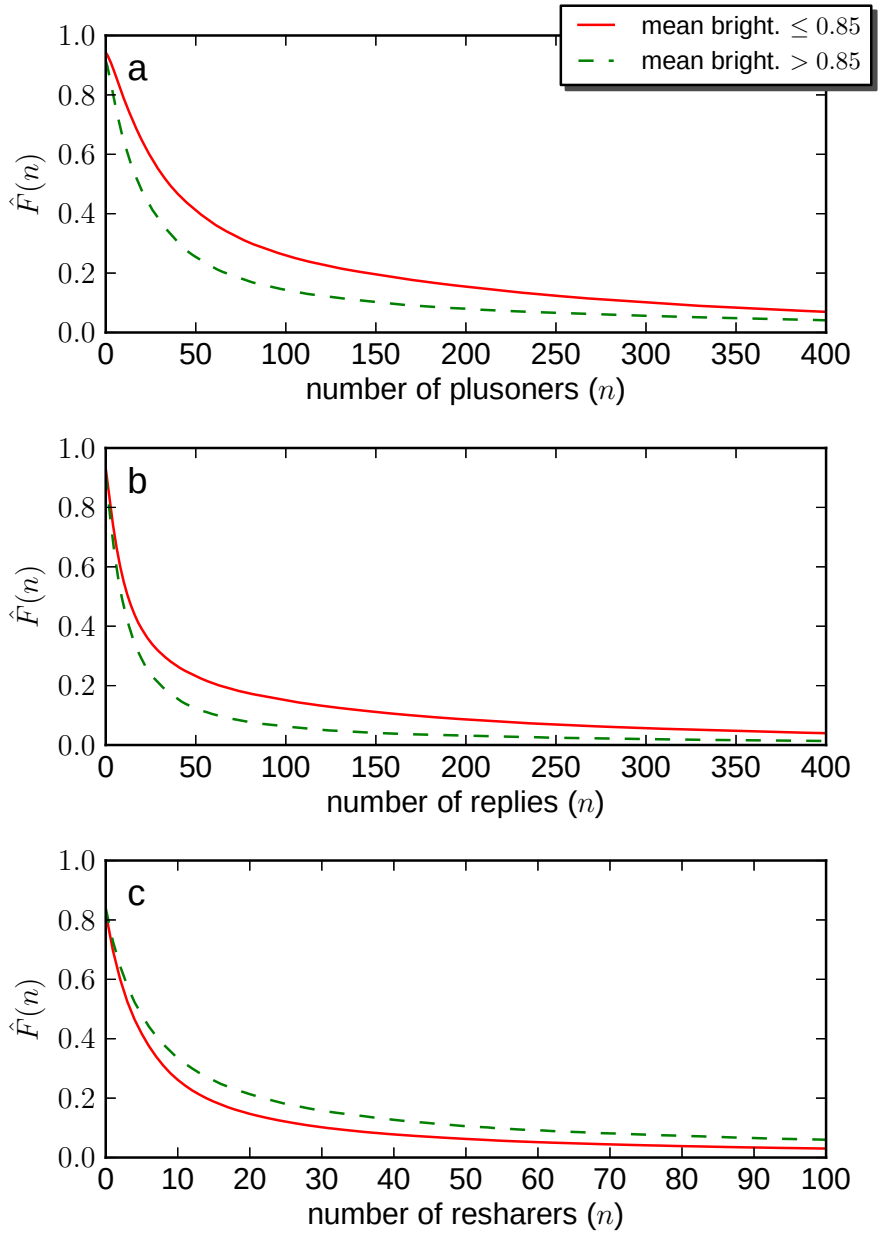


Figure 5.7: Virality CCDFs for image Brightness.

servative threshold of 0.05 (on a 0-1 scale).

As can be seen in Figure 5.8.a and 5.8.b, colored images (with saturation higher than 0.05) have a higher probability of collecting more plusoners and replies as compared to images with lower saturation (grayscale). In particular

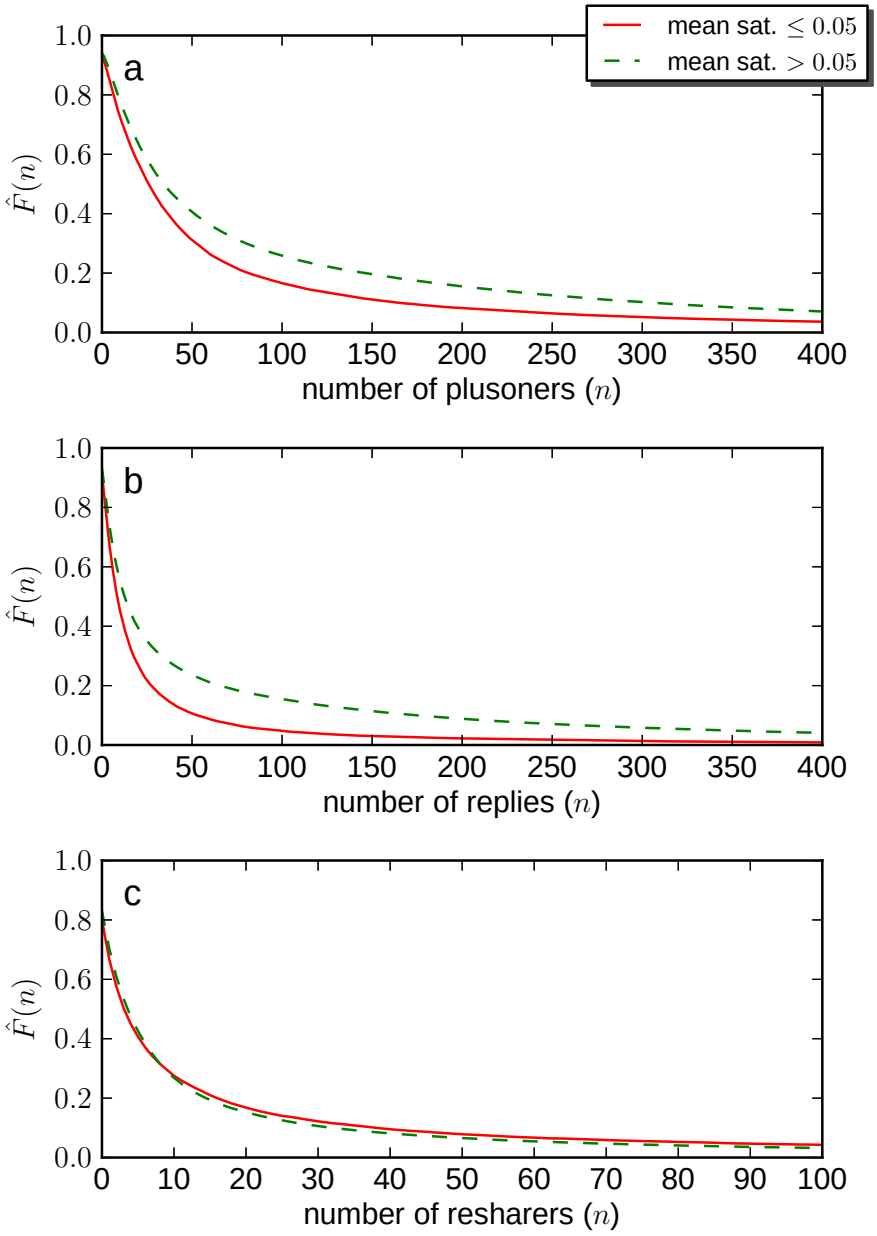


Figure 5.8: Virality CCDFs for Grayscale vs. Colored images.

the probability functions for replies is more than two times higher ($\hat{F}_{repl}(50)$ values are 0.26 vs. 0.10, K–S test $p < 0.001$). Instead, image saturation has no relevant impact on resharers.

Very Bright Images

After converting each image in our dataset to the HSB color space, we extracted its mean Saturation and Brightness. More in detail, the HSB (Hue/Saturation/Brightness) color space describes each pixel in an image as a point on a cylinder: the Hue dimension representing its color within the set of primary-secondary ones, while Saturation and Brightness describe respectively how close to the pure color (i.e. its Hue), and how bright it is.We split the dataset according to images mean brightness using a threshold of 0.85 (in a scale included between 0 and 1). Usually images with such an high mean brightness tend to be cartoon-like images rather than pictures. Previous research [125] has shown that pixel brightness is expected to be higher in cartoon-like (or significantly "photoshopped') than in natural images.

Image brightness level has a strong impact on plusoners and replies, and a milder one on resharers. Brighter images have a lower probability of being viral on the first two indexes (Figure 5.7.a and 5.7.b) and a higher probability on the latter (Figure 5.7.c). In particular, lower brightness images have a plusone and reshare probability almost two times higher ($\hat{F}_{plus}(75)$ = 0.31 vs. 0.18, $\hat{F}_{repl}(50)$ = 0.23 vs. 0.12, K–S test $p < 0.001$), while for resharers it is 27% higher in favor of high brightness images ($\hat{F}_{resh}(10)$ = 0.33 vs. 0.26). Surprisingly, analyzing a small random subsample of 200 very bright images, we found that while 88% of these images contained some text, as we would have expected, only 13% were cartoon/comics and only 13% contained the real picture of an object as subject, even if highly "photoshopped". Above all, only a small amount of these images (21%) was considered funny or memetic[5]. The great majority comprised pictures containing infographics, screenshots of software programs, screenshots of social-networks posts and similar. In this respect we are analyzing a content that is meant to be mainly informative, and is some-

[5]Two annotators were used, four binary categories were provided (contain-text/comics/real-picture-obj/funny). The overall inter-annotator agreement on these categories is high, Cohen's kappa 0.74.

how complementary to the content of animated pictures (mainly intended for amusement, see 5.1.3).

Vertical and Horizontal edges

Finally, we want to report on an explorative investigation we made. We focused on the impact of edges intensity on posts virality. The intensity of vertical/horizontal/diagonal edges was computed using Gaussian filters, based on code used in [284] in the context of real-time visual concept classification. The probability density of the average edges intensity follows a gaussian-like distribution, with mean of about 0.08 (both for horizontal and vertical edges). We divided images into two groups: those having an average edge intensity below the sample mean, and those having an average edge intensity above the mean. Results showed that images with horizontal edge intensity below the sample mean are far more viral on the plusoners and replies indexes, while vertical are less discriminative. Results for horizontal hedges are as follows: $\hat{F}_{plus}(75)$ = 0.36 vs. 0.22, $\hat{F}_{repl}(50)$ = 0.27 vs. 0.14, $\hat{F}_{resh}(10)$ = 0.25 vs. 0.29, K–S test $p < 0.001$. While these results do not have an intuitive explanation, they clearly show that there is room for further investigating the impact of edges.

Virality Indexes Correlation

From the analyses above, virality indexes seem to "move together" (in particular plusoners and replies) while resharers appear to indicate a different phenomenon. We hypothesize that plusoners and replies can be considered as a form of endorsement, while reshares are a form of self-representation. This explains why, for example, pictures containing faces are endorsed but not used for self-representation by VIPs' followers. On the contrary, animated images that usually contain funny material are more likely to provoke reshares for followers' self-representation. In fact, people usually tend to represent themselves with

positive feelings rather than negative ones (especially popular users, see [227]), and positive moods appear to be associated with social interactions [64, 293].

Table 5.2: Virality indexes correlation on the datasets

	Pearson	MIC
Static images		
plusoners vs. replies	0.723	0.433
plusoners vs. resharers	0.550	0.217
replies vs. resharers	0.220	0.126
Animated Images		
plusoners vs. replies	0.702	0.304
plusoners vs. resharers	0.787	0.396
replies vs. resharers	0.554	0.205
Text Only		
plusoners vs. replies	0.802	0.529
plusoners vs. resharers	0.285	0.273
replies vs. resharers	0.172	0.185

This is supported also by the correlation analysis of the three virality indexes, reported in Table 5.2, made on the various datasets we exploited. In this analysis we used both the Pearson coefficient and the recent Maximal Information Coefficient (MIC), considering plusoners ≤ 1200, replies ≤ 400 e resharers ≤ 400. MIC is a measure of dependence introduced in [236] and it is part of the Maximal Information-based Nonparametric Exploration (MINE) family of statistics. MIC is able to capture variable relationships of different nature, penalizing similar levels of noise in the same way. In this study we use the Python package *minepy* [8].

In particular, from Table 5.2 we see that: plusones and replies always have a high correlation while replies and resharers always correlate low. Plusoners and reshares, that have a mild correlation in most cases, correlate highly when it comes to funny pictures, i.e. animated ones. This can be explained by a specific

"procedural" effect: the follower expresses his/her appreciation for the funny picture and, after that, he/she reshares the content. Since resharing implies also writing a comment in the new post, the reply is likely not to be added to the original VIP's post.

In Table 5.3 we sum up the main findings of the paper, comparing the various CCDFs: animated images and infographics have much higher probability of being reshared, while colored images or images containing faces have higher probability of being appreciated or commented. Finally, black and white pictures (grayscale) turn out to be the least "viral" on Google+.

Table 5.3: Summary of main findings of the analysis.

	$\hat{F}_{plus}(75)$	$\hat{F}_{repl}(50)$	$\hat{F}_{resh}(10)$
very bright	0.18	0.12	**0.33**
grayscale	0.21	0.11	0.28
color	**0.31**	**0.24**	0.27
animated	0.17	0.08	**0.48**
one-face $> 10\%$ area	**0.35**	**0.30**	0.23

5.1.4 User Analysis

Finally, we investigate if there is any relevant interaction between images characteristics and VIP's typology. In Table 5.4 we report demographic details[6] on the Google+ dataset, as provided by the users in their profile pages.

In order to investigate possible user category effects in our dataset — that is, if our analyses are also influenced by the type of user posting images rather than by the actual content solely, we evaluated the entropy for each image category

[6] *No Category* denotes users that do not provide any personal information and for which it was not possible to trace back their category; *Not Available* denotes seven accounts that were no more publicly accessible when we gathered demographic info; *Other* denotes very rare and unusual category definitions. The *Neutral* gender refers to pages afferent to "non-humans" like products, brands, websites, firms, etc.

Table 5.4: User demographics in the Google+ dataset.

User-category	**Female (%)**	**Male (%)**	**Neutral (%)**	**Total (%)**
Technology	35 (*19%*)	110 (*61%*)	36 (*20%*)	181 (*19%*)
Photography	41 (*24%*)	130 (*76%*)	1 (*1%*)	172 (*18%*)
Music	96 (*59%*)	48 (*29%*)	19 (*12%*)	163 (*17%*)
Writing	26 (*21%*)	76 (*63%*)	19 (*16%*)	121 (*13%*)
Actor	21 (*36%*)	34 (*59%*)	3 (*5%*)	58 (*6%*)
Entrepreneur	12 (*29%*)	29 (*71%*)	-	41 (*4%*)
Sport	-	22 (*55%*)	18 (*45%*)	40 (*4%*)
Artist	11 (*31%*)	21 (*60%*)	3 (*9%*)	35 (*4%*)
TV	8 (*24%*)	11 (*33%*)	14 (*42%*)	33 (*3%*)
Company	-	-	28 (*100%*)	28 (*3%*)
Website	-	-	23 (*100%*)	23 (*2%*)
Politician	-	19 (*86%*)	3 (*14%*)	22 (*2%*)
No Category	6 (*43%*)	8 (*57%*)	-	14 (*1%*)
Organization	-	-	9 (*100%*)	9 (*1%*)
Not Available	-	-	7 (*100%*)	7 (*1%*)
Other	1 (*33%*)	2 (*67%*)	-	3 (*0%*)
Total	**257 (*27%*)**	**510 (*54%*)**	**183 (*19%*)**	**950 (*100%*)**

over the 16 user categories (as defined in Table 5.4). In Table 5.5 we report the contingency table of image-category entropy distributions over user-categories. Looking at the Kullback-Leibler (KL) divergence of specific image categories with respect to the reference distribution (i.e., taken as the total number of images posted by each user-category), we observe very few but interesting effects due to specific user-categories.

In particular, while all the KL divergences are very small, two of them (for Grayscale and High Brightness, reported in Bold) are an order of magnitude greater than other classes. Interestingly the divergence is explained mainly by the distribution gap in only two User's categories. For High Brightness the gap is mainly given by Technology user category that doubles its probability distribution (from 22% to 40%) and Music and Photography that reduce their probability distribution to one third. This divergence from the reference distribution is consistent with the analysis of the content we made in section 5.1.3: these images where mainly infographics and screenshots of software programs and social networks (so mainly connected to technology). For Grayscale the gap is mainly given by Photography users category that rises by 50% its probability distribution and Music, that reduces it to one third. This gap is consistent with the idea, expressed in Section 5.1.3, that black-and-white photography is a particular form of art expressivity mainly used by professionals.

5.1.5 Conclusions

We have presented a study, based on a novel dataset of Google+ posts, showing that perceptual characteristics of an image can strongly affect the virality of the post embedding it. Considering various kinds of images (e.g. cartoons, panorama or self-portraits) and related features (e.g. orientation, animations) we saw that users' reactions are affected in different ways. We provided a series of analyses to explain the underlying phenomena, using three virality metrics (namely plusoners, replies and resharers). Results suggest that plusoners and

Table 5.5: Contingency table of image-category distributions over user-categories.

User-category	**Grayscale**	**Colored**	**High Brightness**	**Low Brightness**	**w/ Face**	**w/o Face**	**Squared**	**Vertical**	**Horizontal**	**Total**
No Category	7%	6%	9%	6%	5%	7%	4%	5%	7%	6%
Actor	4%	6%	5%	5%	8%	5%	5%	6%	5%	5%
Artist	5%	6%	7%	6%	6%	6%	5%	7%	6%	6%
Company	0%	1%	1%	1%	1%	1%	1%	1%	1%	1%
Entrepreneur	8%	7%	6%	7%	7%	7%	8%	5%	8%	7%
Music	**3%**	16%	**3%**	16%	19%	12%	15%	29%	8%	14%
Not Available	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Organization	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Other	0%	0%	0%	0%	0%	0%	2%	0%	0%	0%
Photography	**31%**	19%	**9%**	22%	15%	23%	23%	14%	23%	20%
Politician	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Sport	0%	3%	1%	3%	4%	2%	2%	2%	3%	2%
Technology	27%	22%	**40%**	20%	19%	24%	16%	18%	25%	22%
TV	1%	2%	1%	2%	3%	1%	5%	1%	2%	2%
Website	1%	2%	2%	2%	2%	2%	1%	1%	2%	2%
Writing	11%	10%	17%	10%	11%	10%	11%	10%	11%	11%
KL-divergence	**0.173**	0.002	**0.259**	0.003	0.027	0.006	0.047	0.076	0.029	

replies “move together” while reshares indicate a distinct users’ reaction. In particular, funny and informative images have much higher probability of being reshared but are associated to different image features (animation and high-brightness respectively), while colored images or images containing faces have higher probability of being appreciated and commented.

Future work will dig deeper into the assessment of relations between visual content and virality indexes, adopting multivariate analysis that includes user’s categories (e.g. which is the viral effect of b/w pictures taken by professional photographer as compared to those taken by non professional users). We will also extend our experimental setup in the following ways: (a) taking into account compositional features of the images, i.e. resembling concepts such as the well-known ”rule of thirds”; (b) extracting and exploiting descriptors such as color histograms, oriented-edges histograms; (c) building upon the vast literature available in the context of scene/object recognition, dividing our dataset into specific categories in order to analyse relations between categories, such as natural images or sport images, and their virality.

5.2 A Lexicon for Emotion Analysis from Crowd-Annotated News

Sentiment analysis has proved useful in several application scenarios, for instance in buzz monitoring – the marketing technique for keeping track of consumer responses to services and products – where identifying positive and negative customer experiences helps to assess product and service demand, tackle crisis management, etc.

Instead, the use of finer-grained models, accounting for the role of individual emotions, is still in its infancy. Still, the simple division in ‘positive’ vs. ‘negative’ comments may not suffice, as in these examples: ‘*I’m so miserable, I dropped my IPhone in the water and now it’s not working anymore*’ (SADNESS)

vs. '*I am very upset, my new IPhone keeps not working!*' (ANGER). While both texts express a negative sentiment, only the latter, connected to anger, is relevant for buzz monitoring. Thus, emotion analysis represents a natural evolution of sentiment analysis.

Many approaches to sentiment analysis make use of lexical resources – i.e. lists of positive and negative words – often deployed as baselines or as features for other methods, usually machine learning based [175]. In these lexica, words are associated with their prior polarity, i.e. whether such word out of context evokes something positive or something negative. For example, *wonderful* has a positive connotation – prior polarity – while *horrible* has a negative one.

The quest for a high precision and high coverage lexicon, where words are associated with either sentiment or emotion scores, has several reasons. First, it is fundamental for tasks such as affective modification of existing texts, where words' polarity together with their score are necessary for creating multiple *graded* variations of the original text [102, 127, 305].

Second, considering words order makes a difference in sentiment analysis. This calls for a role of compositionality, where the score of a sentence is computed by composing the scores of the words up in the syntactic tree. Works worth mentioning in this connection are: [251], that uses recursive neural networks to learn compositional rules for sentiment analysis, and [200, 202] that exploit hand-coded rules to compose the emotions expressed by words in a sentence. In this respect, compositional approaches represent a new promising trend, since all other approaches, either using semantic similarity or Bag-of-Words (BOW) based machine-learning, cannot handle, for example, cases of texts with same wording but different words order: "*The dangerous killer escaped one month ago, but lately he was arrested*" (RELIEF, HAPPYNESS) vs. "*The dangerous killer was arrested one month ago, but lately he escaped*" (FEAR). The work in [296] partially accounts for this problem and argues that using word bigram features allows improving over BOW based methods, where

words are taken as features in isolation. This way it is possible to capture simple compositional phenomena like polarity reversing in "*killing cancer*".

Finally, tasks such as copywriting, where evocative names are a key element to a successful product [211, 212] require exhaustive lists of emotion related words. In such cases no context is given and the brand name alone, with its perceived prior polarity, is responsible for stating the area of competition and evoking semantic associations. For example *Mitsubishi* changed the name of one of its SUV for the Spanish market, since the original name *Pajero* had a very negative prior polarity, as it means 'wanker' in Spanish [224]. Evoking emotions in also fundamental for a successful name: consider names of a perfume like *Obsession*, or technological products like MacBook *air*.

In this work, we aim at automatically producing a high coverage and high precision emotion lexicon using distributional semantics, with numerical scores associated with each emotion, like it has already been done for sentiment analysis. To this end, we take advantage in an original way of massive crowdsourced affective annotations associated with news articles, obtained by crawling the `rappler.com` social news network. We also evaluate our lexicon by integrating it in unsupervised classification and regression settings for emotion recognition. Results indicate that the use of our resource, even if automatically acquired, is highly beneficial in the affective text recognition scenario.

5.2.1 Related Work

Within the broad field of sentiment analysis, we hereby provide a short review of research efforts put towards building sentiment and emotion lexica, regardless of the approach in which such lists are then used (machine learning, rule based or deep learning). A general overview can be found in [175, 213, 214, 310].

Sentiment Lexica. In recent years there has been an increasing focus on producing lists of words (lexica) with prior polarities, to be used in sentiment

	AFRAID	AMUSED	ANGRY	ANNOYED	DONT_CARE	HAPPY	INSPIRED	SAD
doc_10002	0.75	0.00	0.00	0.00	0.00	0.00	0.25	0.00
doc_10003	0.00	0.50	0.00	0.16	0.17	0.17	0.00	0.00
doc_10004	0.52	0.02	0.03	0.02	0.02	0.06	0.02	0.31
doc_10011	0.40	0.00	0.00	0.20	0.00	0.20	0.20	0.00
doc_10028	0.00	0.30	0.08	0.00	0.00	0.23	0.31	0.08

Table 5.6: An excerpt of the Document-by-Emotion Matrix - M_{DE}

analysis. When building such lists, a trade-off between coverage of the resource and its precision is to be found.

One of the most well-known resources is *SentiWordNet* (SWN) [81, 20], in which each entry is a set of `lemma#PoS#sense-number` sharing the same meaning, called *synset*. Each synset `s` is associated with the numerical scores `Pos(s)` and `Neg(s)`, ranging from 0 to 1. These scores – automatically assigned starting from a bunch of seed terms – represent the positive and negative valence (or posterior polarity) of the synset and are inherited by each `lemma#PoS#sense-number` in the synset.

Starting from SWN, several prior polarities for words (*SWN-prior*), in the form `lemma#PoS`, can be computed (e.g. considering only the first-sense, averaging on all the senses, etc.). These approaches, detailed in [104], produce a list of 155k words, where the lower precision given by the automatic scoring of SWN is compensated by the high coverage.

Another widely used resource is *ANEW* [31], providing valence scores for 1k words, which were manually assigned by several annotators. This resource has a low coverage, but the precision is maximum. Similarly, the *SO-CAL* entries [272] were manually tagged by a small number of annotators with a multi-class label (from `very_negative` to `very_positive`). These ratings were further validated through crowd-sourcing, ending up with a list of roughly 4k words. More recently, a resource that replicated ANEW annotation approach using crowd-sourcing, was released [298], providing sentiment scores for 14k

words. Interestingly, this resource annotates the most frequent words in English, so, even if lexicon coverage is still far lower than SWN-prior, it grants a high coverage, with human precision, of language use.

Finally, the *General Inquirer* lexicon [262] provides a binary classification (`positive`/`negative`) of 4k sentiment-bearing words, while the resource in [309] expands the General Inquirer to 6k words.

Emotion Lexica. Compared to sentiment lexica, far less emotion lexica have been produced, and all have lower coverage. One of the most used resources is *WordNetAffect* [263] which contains manually assigned affective labels to WordNet synsets (ANGER, JOY, FEAR, etc.). It currently provides 900 annotated synsets and 1.6k words in the form `lemma#PoS#sense`, corresponding to roughly 1 thousand `lemma#PoS`.

AffectNet, part of the SenticNet project [38], contains 10k words (out of 23k entries) taken from ConceptNet and aligned with WordNetAffect. This resource extends WordNetAffect labels to concepts like 'have breakfast'. *Fuzzy Affect Lexicon* [266] contains roughly 4k `lemma#PoS` manually annotated by one linguist using 80 emotion labels. Finally *Affect database* is an extension of SentiFul [201] and contains 2.5K words in the form `lemma#PoS`. The latter is the only lexicon providing words annotated also with emotion scores rather than only with labels.

5.2.2 Dataset Collection

To build our emotion lexicon we harvested all the news articles from `rappler.com`, as of June 3rd 2013: the final dataset consists of 13.5 M words over 25.3 K documents, with an average of 530 words per document. For each document, along with the text we also harvested the information displayed by Rappler's *Mood Meter*, a small interface offering the readers the opportunity to click on the emotion that a given Rappler story made them feel. The idea behind the

Mood Meter is actually "getting people to *crowdsource* the mood for the day"[7], and returning the percentage of votes for each emotion label for a given story. This way, hundreds of thousands votes have been collected since the launch of the service. In our novel approach to 'crowdsourcing', as compared to other NLP tasks that rely on tools like Amazon's Mechanical Turk [250], the subjects are aware of the 'implicit annotation task' but aware of the 'annotation task' but they are not paid. From this data, we built a document-by-emotion matrix M_{DE}, providing the voting percentages for each document in the eight affective dimensions available in Rappler. An excerpt is provided in Table 5.6.

The idea of using documents from the Web annotated with emotions is not new [193, 265], but these works had the limitations of providing a single emotion label per document, rather than a score for each emotion, and, moreover, the annotation was performed by the author of the document alone.

Table 5.7 reports the average percentage of votes for each emotion on the whole corpus: HAPPINESS has a far higher percentage of votes (at least three times). There are several possible explanations, out of the scope of the present paper, for this bias: (i) it is due to cultural characteristics of the audience (Rappler is a Philippine based social news network); (ii) the bias is in the dataset itself, being formed mainly by 'positive' news; (iii) it is a psychological phenomenon due to the fact that people tend to express more positive moods on social networks [64, 227, 293]. In any case, the predominance of happy mood has been found in other datasets, for instance `LiveJournal.com` posts [265].

In the following section we will discuss how we handled this problem.

5.2.3 Emotion Lexicon Creation

As a next step we built a word-by-emotion matrix starting from M_{DE} using an approach based on compositional semantics. To do so, we first lemmatized and

[7] http://www.niemanlab.org/2012/08/in-the-philippines-rappler-is-trying-to-figure-out-the-role-of-emotion-in-the-news/

EMOTION	Votes$_\mu$	EMOTION	Votes$_\mu$
AFRAID	0.04	DONT_CARE	0.05
AMUSED	0.10	HAPPY	0.32
ANGRY	0.10	INSPIRED	0.10
ANNOYED	0.06	SAD	0.11

Table 5.7: Average percentages of votes.

Word	AFRAID	AMUSED	ANGRY	ANNOYED	DONT_CARE	HAPPY	INSPIRED	SAD
awe#n	0.08	0.12	0.04	0.11	0.07	0.15	**0.38**	0.05
comical#a	0.02	**0.51**	0.04	0.05	0.12	0.17	0.03	0.06
crime#n	0.11	0.10	**0.23**	0.15	0.07	0.09	0.09	0.15
criminal#a	0.12	0.10	**0.25**	0.14	0.10	0.11	0.07	0.11
dead#a	0.17	0.07	0.17	0.07	0.07	0.05	0.05	**0.35**
funny#a	0.04	**0.29**	0.04	0.11	0.16	0.13	0.15	0.08
future#n	0.09	0.12	0.09	0.12	0.13	0.13	**0.21**	0.10
game#n	0.06	0.15	0.06	0.08	0.15	**0.23**	0.15	0.12
kill#v	**0.23**	0.06	**0.21**	0.07	0.05	0.06	0.05	**0.27**
rapist#n	0.02	0.07	**0.46**	0.07	0.08	0.16	0.03	0.12
sad#a	0.06	0.12	0.09	0.14	0.13	0.07	0.15	**0.24**
warning#n	**0.44**	0.06	0.09	0.09	0.06	0.06	0.04	0.16

Table 5.8: An excerpt of the Word-by-Emotion Matrix (M_{WE}) using normalized frequencies (nf). Emotions weighting more than 20% in a word are highlighted for readability purposes.

PoS tagged all the documents (where PoS can be adj., nouns, verbs, adv.) and kept only those `lemma#PoS` present also in WordNet, similar to SWN-prior and WordNetAffect resources.

We then computed the term-by-document matrices using raw frequencies, normalized frequencies, and tf-idf ($M_{WD,f}$, $M_{WD,nf}$ and $M_{WD,tfidf}$ respectively), so to test which of the three weights is better. After that, we applied matrix multiplication between the document-by-emotion and word-by-document matrices ($M_{DE} \bullet M_{WD}$) to obtain a (raw) word-by-emotion matrix M_{WE}. This method allows us to 'merge' words with emotions by summing the products of the weight of a word with the weight of the emotions in each document.

Finally, we transformed M_{WE} by first applying normalization column-wise

(so to eliminate the over representation for happiness as discussed in Section 5.2.2) and then scaling the data row-wise so to sum up to one. An excerpt of the final Matrix M_{WE} is presented in Table 5.8, and it can be interpreted as a list of words with scores that represent how much weight a given word has in the affective dimensions we consider. So, for example, `awe#n` has a predominant weight in INSPIRED (0.38), `comical#a` has a predominant weight in AMUSED (0.51), while `kill#v` has a predominant weight in AFRAID, ANGRY and SAD (0.23, 0.21 and 0.27 respectively). This matrix, that we call `DepecheMood`[8], represents our emotion lexicon, it contains 37k entries and is freely available for research purposes at [anonymous-link].

5.2.4 Experiments

To evaluate the performance we can obtain with our lexicon, we use the public dataset provided for the SemEval 2007 task on 'Affective Text' [264]. The task was focused on emotion recognition in one thousand news headlines, both in regression and classification settings. Headlines typically consist of a few words and are often written with the intention to 'provoke' emotions so to attract the readers' attention. An example of headline from the dataset is the following: "*Iraq car bombings kill 22 People, wound more than 60*". For the regression task the values provided are: `<anger(0.32),disgust(0.27),fear(0.84), joy(0.0),sadness(0.95),surprise(0.20)>` while for the classification task the labels provided are {`FEAR,SADNESS`}. This dataset is of interest to us since the 'compositional' problem is less prominent given the simplified syntax of news headlines, containing, for example, fewer adverbs (like negations or intensifiers) than normal sentences [283]. adverbs Furthermore, this is to our knowledge the only dataset available providing numerical scores for emotions. Finally, this dataset was meant for unsupervised approaches (just a small trial sample was provided), so to avoid simple text categorization approaches.

[8] In french 'depeche' means dispatch/news.

As the affective dimensions present in the test set – based on the six basic emotions model [78] – do not exactly match with the ones provided by Rappler's Mood Meter, we first define a mapping between the two, reported in Table 5.9. Then, we proceed to transform the test headlines to the `lemma#PoS` format.

SemEval	Rappler	SemEval	Rappler
FEAR	AFRAID	SURPRISE	**INSPIRED**
ANGER	ANGRY	DISGUST	**ANNOYED**
JOY	HAPPY	-	AMUSED
SADNESS	SAD	-	DON'T CARE

Table 5.9: Mapping of labels between our dataset and Semeval2007. In bold, labels that do not have a precise semantic mapping.

Only one test headline contained exclusively words not present in `DepecheMood`, further indicating the high-coverage nature of our resource. In Table 5.10 we report the coverage of some Sentiment and Emotion Lexica of different sizes on the same dataset. Similar to Warriner et. al (2013), we observe that even if the number of entries of our lexicon is far lower than SWN-prior approaches, the fact that we extracted and annotated words from documents grants a high coverage of language use.

Sentiment Lexica	ANEW	1k entries	0.10
	Warriner et. al	13k entries	0.51
	SWN-prior	155k entries	**0.67**
Emotion Lexica	WNAffect	1k entries	0.12
	DepecheMood	37k entries	**0.64**

Table 5.10: Statistics on words coverage per headline.

Since our primary goal is to assess the quality of `DepecheMood` we first focus on the regression task. We do so by using a very naïve approach, similar to "WordNetAffect presence" discussed in [265]: for each headline, we simply compute a value, for any affective dimension, by averaging the corresponding

affective scores –obtained from DepecheMood- of all lemma#PoS present in the headline.

In Table 5.11 we report the results obtained using the three versions of our resource (Pearson correlation), along with the best performance on each emotion of other systems[9] ($best_{se}$); the last column contains the upper bound of inter-annotator agreement. For 5 emotions out of 6 we improve over the best performing systems. We do not outperform other systems only on DISGUST, that has no clear alignment with our labels (see Table 5.9).

Interestingly, even using a sub-optimal alignment for SURPRISE we still manage to outperform other systems. Considering the naïve approach we used, we can reasonably conclude that the quality and coverage of our resource are the reason of such results, and that adopting more complex approaches can possibly further improve performances in text-based emotion recognition.

	DepecheMood			$best_{se}$	upper
	f	*nf*	*tfidf*		
FEAR	**0.56**	0.54	0.53	0.45	0.64
ANGER	0.36	**0.38**	0.36	0.32	0.50
SURPRISE*	**0.25**	0.21	0.24	0.16	0.36
DISGUST*	0.05	0.06	0.07	**0.18**	0.44
JOY	0.39	**0.40**	0.39	0.26	0.60
SADNESS	**0.48**	0.47	0.46	0.41	0.68

Table 5.11: Regression results – Pearson's correlation

As a final test, we evaluate our resource in the classification task. The naïve approach used in this case consists in mapping the average of the scores of all words in the headline to a binary decision with fixed threshold at 0.5 for each emotion (after min-max normalization on all test headlines scores). In Table 5.12 we report the results (F1 measure) of our approach along with the best performance of other systems on each emotion ($best_{se}$), as in the previous case. For 3 emotions out of 6 we improve over the best performing systems,

[9]Systems participating to the 'Affective Text' task plus the approaches presented in [265].

for one emotion we manage to obtain the same results, and for 2 emotions we do not outperform other systems. In this case the difference in performances among the various ways of representing the word-by-document matrix is more prominent: raw frequencies (f) provide the worst results.

	$DepecheMood$			$best_{se}$
	f	*nf*	*tfidf*	
FEAR	0.25	**0.32**	0.31	0.23
ANGER	0.00	0.00	0.00	**0.17**
SURPRISE*	0.13	**0.16**	0.09	0.15
DISGUST*	0.03	0.02	0.01	**0.04**
JOY	0.22	0.30	**0.32**	**0.32**
SADNESS	0.36	**0.40**	0.38	0.30

Table 5.12: Classification results – F1 measures

5.2.5 Conclusions

We presented and provided to the community `DepecheMood`, an emotion lexicon built in a novel and totally automated way by harvesting crowd-sourced affective annotation from a social news network. Our experimental results indicate high-coverage and high-precision of the lexicon, showing significant improvements over state-of-the-art unsupervised approaches even when using the resource with very naïve classification and regression strategies. We believe that the wealth of information provided by social media can be harnessed to build models and resources for emotion recognition from text, going a step beyond sentiment analysis. Our future work will include testing Singular Value Decomposition on the word-by-document matrices, allowing to propagate emotions values for a document to similar words non present in the document itself, and the study of perceived mood effects on virality indices and readers engagement by exploiting tweets, likes, reshares and comments.

Chapter 6

Conclusions

In this thesis, we have addressed the problem of automatic behavior understanding from a wide range of perspectives.

In Chapter 2, we have tackled a scenario in which subjects interact with technology or enjoy multimedia contents; insights from the presented works can be valuable to design and implement cheaper and less obtrusive evaluation strategies, and to provide more engaging experiences to viewers (especially, with the current deployment of so-called *Smart TVs*.

In Chapter 3, we moved to *small group* scenarios, with the design, development, and evaluation of integrated software systems able to capture interaction patterns, such as *social attention*, and to infer high level behavioral determinants, such as *personality*. In particular, we have contributed to the research community the very first computational approaches to the detection of *personality states* [256].

In Chapter 4 we further loosened constraints, exploiting wearable and mobile devices to extract and store quantitative behavioral manifestations, along with questionnaire sampling to assess their validity. We showed how metadata can be valuable to infer personality [255], and provided an important contribution with the SocioMetric Badges Corpus, so far the largest database available for Organizational Behavior studies.

Finally, in Chapter 5 we focused on scenarios in which we have absolutely no control: crawling the *world wide web* in order to extract higher level knowledge from behavioral traces left online. For example, we showed the relation between image characteristics and virality on a popular social network [105]; furthermore, we harvested a popular news website which provided its readers with a simple tool to express their emotional feedback: using such data, we were able to build, and contribute to the community, a high-coverage and high-precision emotional lexicon.

We live interesting times.

In 1916, the great Italian thinker Antonio Gramsci wrote *Facts ripen in the dark, for masters sew the fabric of public life unbeknownst to puppets.* [1](translation by the author).

A century later, we find ourselves living in a world vastly reshaped by the continuous development and availability of information, and the related technological means of production and distribution. Still, policies and regulations move at a significantly lower pace, mostly trying to update existing and outdated legal frameworks with the effect of keeping the relations of power mostly intact.

Nonetheless, the paradigm-shift provided by the data-driven society whose early days we are now witnessing seems ineluctable. Even the most complex and pervasive surveillance system was brought to light and is under public scrutiny, thanks to data and information that spread, reached and enraged billions of citizens. The very technology we are working on, as any technology, can be used for advancing humanity or for humiliating it.

Previous research works [203, 209, 226, 233] have shown how cooperation is as important and prevalent in human society as competition. The *mathematical, predictive science of society that includes both individual differences and*

[1]"Dei fatti maturano nell'ombra, perché mani non sorvegliate da nessun controllo tessono la tela della vita collettiva, e la massa ignora." Antonio Gramsci, 1916. From *Sotto la mole*, Einaudi, Torino, 1964, p. 228.

the relationships between individuals envisioned by Alex "Sandy" Pentland in his latest book *Social Physics* [218] and under current scientific development, aims in fact to *engineer better social systems*, building a feedback loop between scientific knowledge and *Big Data* and allowing distributed realtime decision making in a decentralized fashion.

We already often use new services and innovations spawned from the *sharing economy* (e.g. AirBnb, ZipCar, etc.): *we're moving from a world where we're organized around ownership to one organized around access to assets*[2]. Still, as mentioned in the introduction of this thesis, as personal data becomes an asset, service providers (the asset owners) tend naturally to keep such information in *silos*, deriving value from it in old-fashioned ways, often with little or no return for its actual producers (*us*). An economic paradigm based on *liberation* of data from such silos has recently been proposed by Doc Searls: the *intention economy* [246].

The complete transition from old-fashioned market-based economies to exchange- or intention- based economic paradigms will be possible if we finally stop thinking about data with 1970s mindset: as John Clippinger put it, elaborating the concept of Social Stack[3] within the ID3 initiative, *People could be empowered to control their personal information and identities, and to develop bonds of social trust and reputation in stable, enduring online communities*. We are starting to look at data as *water* (a common asset) rather than as *oil* (a commodity), and imagining new forms of governance and self-organizing communities built on data and decentralized services.

Indeed, in the quest for understanding the potential for a user-centric perspective, as opposed to market-centric, on mobile personal data, we have presented in Section 4.4 a case study on the economics of personal mobile data.

It is our understanding that revolutions happen through cooperation, and are

[2]Lisa Gansky, http://www.forbes.com/sites/tomiogeron/2013/01/23/airbnb-and-the-unstoppable-rise-of-the-share-economy/

[3]http://idcubed.org/open-platform/socialstack/

failed by competition.

Bibliography

[1] Jie Zhang Aaditeshwar Seth and Robin Cohen. A multi-disciplinary approach for recommending weblog messages. In *The AAAI 2008 Workshop on Enhanced Messaging*, 2008.

[2] Mojtaba Khomami Abadi, Jacopo Staiano, Alessandro Cappelletti, Massimo Zancanaro, and Nicu Sebe. Multimodal engagement classification for affective cinema. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 411–416. IEEE, 2013.

[3] Alessandro Acquisti and Jens Grossklags. Privacy and rationality in individual decision making. *IEEE Security & Privacy*, 2:24–30, 2005.

[4] Alessandro Acquisti, Leslie John, and George Loewenstein. What is privacy worth. In *Twenty first workshop on information systems and economics (WISE)*, pages 14–15, 2009.

[5] E. Adar and B. Huberman. A market for secrets. *First Monday*, 6(8), 2001.

[6] N Aharony, W Pan, C Ip, I Khayal, and A Pentland. Social fmri: Investigating and shaping social mechanisms in the real world. *Pervasive Mob. Comput*, 7(6):643–659, 2011.

[7] Yong-Yeol Ahn, Seungyeop Han, Haewoon Kwak, Sue Moon, and Hawoong Jeong. Analysis of topological characteristics of huge online so-

cial networking services. In *Proceedings of the 16th international conference on World Wide Web*, pages 835–844. ACM, 2007.

[8] Davide Albanese, Michele Filosi, Roberto Visintainer, Samantha Riccadonna, Giuseppe Jurman, and Cesare Furlanello. minerva and minepy: a c engine for the mine suite and its r, python and matlab wrappers. *Bioinformatics*, 29(3):407–408, 2013.

[9] N. Ambady and R. Rosenthal. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological bulletin*, 111(2):256, 1992.

[10] Yair Amichai-Hamburger and Gideon Vinitzky. Social network use and personality. *Computers in Human Behavior*, 26(6):1289–1295, 2010.

[11] Christina Aperjis and Bernardo A Huberman. A market for unbiased private data: Paying individuals according to their privacy attitudes. *arXiv preprint arXiv:1205.0030*, 2012.

[12] Sinan Aral, Erik Brynjolfsson, and Marshall Van Alstyne. Information, Technology and Information Worker Productivity: Task Level Evidence. *NBER Working Paper*, 13172:1–32, 2007.

[13] Sinan Aral and Dylan Walker. Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Management Science*, 57(9):1623–1639, 2011.

[14] Shlomo Argamon, Sushant Dhawle, Moshe Koppel, and James Pennebaker. Lexical predictors of personality type. 2005.

[15] D. Ariely and Z. Carmon. Gestalt characteristics of experiences: the defining feature of summarized events. *Journal of Behavioural Decision Making*, 13(2):191–201, 2000.

[16] S. Arifin and P.Y.K. Cheung. A computation method for video segmentation utilizing the pleasure-arousal-dominance emotional information. In *ACM International Conference on Multimedia*, 2007.

[17] Rudolf Arnheim. *Art and visual perception*. Stockholms Universitet, 1987.

[18] Sileye O. Ba and Jean-Marc Odobez. Recognizing visual focus of attention from head pose in natural meetings. *Trans. Sys. Man Cyber. Part B*, 39(1):16–33, 2009.

[19] Silye O. Ba and Jean-Marc Odobez. Visual Focus of Attention Estimation from Head Pose Posterior Probability Distributions. In *ICME*, 2008.

[20] S. Baccianella, A. Esuli, and F. Sebastiani. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC '10)*, pages 2200–2204, Valletta, Malta, 2010.

[21] Louise Barkhuus, Barry Brown, Marek Bell, Scott Sherwood, Malcolm Hall, and Matthew Chalmers. From awareness to repartee: sharing location within social groups. In *proceedings of the SIGCHI conference on human factors in computing systems*, pages 497–506. ACM, 2008.

[22] Murray R. Barrick, Michael K. Mount, and Timothy A. Judge. Personality and Performance at the Beginning of the New Millennium: What Do We Know and Where Do We Go Next? *International Journal of Selection and Assessment*, 9(1&2):9–30, 2001.

[23] S. Barsade and Brief A. *The affective revolution in organizational behavior: The emergence of a paradigm*, pages 3–52. Londin:Lawrence Erlbaum Associates, Publishers, 2 edition, 2003.

[24] Sigal G Barsade. The Ripple Effect: Emotional Contagion and Its Influence on Group Behavior. *Administrative Science Quarterly*, 47(4):644, 2002.

[25] Ligia Maria Batrinca, Nadia Mana, Bruno Lepri, Fabio Pianesi, and Nicu Sebe. Please, tell me about yourself: automatic personality assessment using short self-presentations. In *Proceedings of the 13th international conference on multimodal interfaces*, ICMI '11, pages 255–262, New York, NY, USA, 2011. ACM.

[26] Joey Benedek and Trish Miner. Measuring desirability: New methods for evaluating desirability in a usability lab setting. *Proceedings of Usability Professionals Association*, pages 8–12, 2002.

[27] Jonah A. Berger and Katherine L. Milkman. Social Transmission, Emotion, and the Virality of Online Content. *Social Science Research Network Working Paper Series*, December 2009.

[28] Phillip Bonacich. Power and centrality: A family of measures. *American Journal of Sociology*, 92(5):1170–1182, 1987.

[29] Stephen P. Borgatti and Pacey C. Foster. The network paradigm in organizational research: A review and typology. *Journal of Management*, 29(2):991–1013, 2003.

[30] Margaret M. Bradley and Peter J. Lang. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49–59, 1994.

[31] M.M. Bradley and P.J. Lang. Affective norms for english words (anew): Instruction manual and affective ratings. *Technical Report C-1, University of Florida*, 1999.

[32] Pedro Branco, Peter Firth, L. Miguel Encarnação, and Paolo Bonato. Faces of emotion in human-computer interaction. In *CHI '05 extended abstracts on Human factors in computing systems*, CHI EA '05, pages 1236–1239, New York, NY, USA, 2005. ACM.

[33] Leo Breiman. Random forests. In *Machine Learning*, pages 5–32, 2001.

[34] Willem-Paul Brinkman and Nick Fine. Towards customized emotional design: an explorative study of user personality and user interface skin preferences. In *Proceedings of the 2005 annual conference on European association of cognitive ergonomics*, EACE '05, pages 107–114, 2005.

[35] Brandon Burr. Vaca: a tool for qualitative video analysis. In *CHI '06 extended abstracts on Human factors in computing systems*, CHI EA '06, pages 622–627, New York, NY, USA, 2006. ACM.

[36] Ronald S Burt. *Structural Holes: The Social Structure of Competition*, volume 5. Harvard University Press, 1992.

[37] R. Calvo and S. D'Mello. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, 1(1):18–37, 2010.

[38] Erik Cambria and Amir Hussain. *Sentic computing*. Springer, 2012.

[39] Juan Pablo Carrascal, Christopher Riederer, Vijay Erramilli, Mauro Cherubini, and Rodrigo de Oliveira. Your browsing behavior for a big mac: Economics of personal information online. In *Proceedings of the 22nd international conference on World Wide Web*, pages 189–200, 2013.

[40] Rich Caruana, Nikos Karampatziakis, and Ainur Yessenalina. An empirical evaluation of supervised learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning*, ICML '08, pages 96–103, 2008.

[41] Ching Hau Chan and Gareth J. F. Jones. Affect-based indexing and retrieval of films. In *ACM International Conference on Multimedia*, pages 427–430, 2005.

[42] L.S. Chen. *Joint Processing of Audio-visual Information for the Recognition of Emotional Expressions in Human-computer Interaction*. PhD thesis, University of Illinois at Urbana-Champaign, 2000.

[43] Y. Cheon and D. Kim. Natural facial expression recognition using differential-AAM and manifold learning. *Pattern Recognition*, 42(7):260–274, 2009.

[44] Mauro Cherubini and Nuria Oliver. A refined experience sampling method to capture mobile user experience. *CoRR*, abs/0906.4125, 2009.

[45] Gokul Chittaranjan, Jan Blom, and Daniel Gatica-Perez. Who's who with big-five: Analyzing and classifying personality. In *ISWC 2011, the fifteenth annual International Symposium on Wearable Computers*, 2011.

[46] Gokul Chittaranjan, Jan Blom, and Daniel Gatica-Perez. Mining large-scale smartphone data for personality studies. *Personal and Ubiquitous Computing*, 17(3):433–450, 2013.

[47] Konstantinos Chorianopoulos and Diomidis Spinellis. User interface evaluation of interactive tv: a media studies perspective. *Univers. Access Inf. Soc.*, 5(2):209–218, 2006.

[48] Emily Christofides, Amy Muise, and Serge Desmarais. Hey mom, whats on your facebook? comparing facebook disclosure and privacy in adolescents and adults. *Social Psychological and Personality Science*, 3(1):48–54, 2012.

[49] David Cohen and B. Neil Cuffin. Demonstration of useful differences between magnetoencephalogram and electroencephalogram. *Electroencephalography and clinical neurophysiology*, 56(1):38–51, 07 1983.

[50] I. Cohen, N. Sebe, L. Chen, A. Garg, and T.S. Huang. Facial expression recognition from video sequences: Temporal and static modeling. *Computer Vision and Image Understanding*, 91(1–2):160–187, 2003.

[51] I. Cohen, N. Sebe, F. Cozman, M. Cirelo, and T.S. Huang. Semi-supervised learning of classifiers: Theory, algorithms, and applications to human-computer interaction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(12):1553–1567, 2004.

[52] Sunny Consolvo, Ian E Smith, Tara Matthews, Anthony LaMarca, Jason Tabert, and Pauline Powledge. Location disclosure to social relations: why, when, & what people want to share. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 81–90. ACM, 2005.

[53] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[54] Paul T Costa and Robert R McCrae. Professional manual: revised neo personality inventory (neo-pi-r) and neo five-factor inventory (neo-ffi). *Odessa, FL: Psychological Assessment Resources*, 1992.

[55] P.T. Costa, R.R. McCrae, and I.C. Siegler. *Continuity and change over the adult life cycle: Personality and personality disorders*, pages 129–153. 1999.

[56] P.C. Cozby. Self-disclosure: A literature review. *Psychological Bulletin*, 79(2):73–91, 1973.

[57] J R Crawford and J D Henry. The positive and negative affect schedule (panas): Construct validity, measurement properties and normative data in a large non-clinical sample. *British Journal of Clinical Psychology*, 43:245–265, 2004.

[58] Jared R Curhan and Alex Pentland. Thin slices of negotiation: predicting outcomes from conversational dynamics within the first 5 minutes. *Journal of Applied Psychology*, 92(3):802, 2007.

[59] Dan Cvrcek, Marek Kumpost, Vashek Matyas, and George Danezis. A study on the value of location privacy. In *Proceedings of the 5th ACM workshop on Privacy in electronic society*, pages 109–118. ACM, 2006.

[60] C. Danescu-Niculescu-Mizil, J. Cheng, J. Kleinberg, and L. Lee. You had me at hello: How phrasing affects memorability. In *Proceedings of the ACL*, 2012.

[61] George Danezis, Stephen Lewis, and Ross J Anderson. How much is location privacy worth? In *WEIS*, volume 5. Citeseer, 2005.

[62] R. J. Davidson, P. Ekman, C. Saron, J. Senulis, and W. V. Friesen. Approach/withdrawal and cerebral asymmetry: Emotional expression and brain physiology. *Journal of Personality and Social Psychology*, 58(2):330–341, 1990.

[63] James A Davis and Samuel Leinhardt. The structure of positive interpersonal relations in small groups. *Sociological Theories in Progress*, pages 218–251, 1972.

[64] Munmun De Choudhury, Scott Counts, and Michael Gamon. Not all moods are created equal! exploring human emotional states in social media. In *Proceedings of ICWSM-12*, 2012.

[65] Yves-Alexandre de Montjoye, Jordi Quoidbach, Florent Robic, and Alex Pentland. Predicting personality using novel mobile phone-based metrics. In *SBP*, pages 48–55, 2013.

[66] Yves-Alexandre de Montjoye, Samuel S Wang, Alex Pentland, Dinh Tien Tuan Anh, Anwitaman Datta, et al. On the trusted use of large-scale personal data. *IEEE Data Eng. Bull.*, 35(4):5–8, 2012.

[67] M Den Uyl, H Van Kuilenburg, and E Lebert. *FaceReader: an online facial expression recognition system*. 2005.

[68] Christian M. Derbaix. The impact of affective reactions on attitudes toward the advertisement and the brand: A step toward ecological validity. *Journal of Marketing Research*, 32(4):470–479, 1995.

[69] Pieter Desmet. Funology. chapter Measuring emotion: development and application of an instrument to measure emotional responses to products, pages 111–123. Kluwer Academic Publishers, 2004.

[70] L Di Blas and M Perugini. L'approccio lessicale nella lingua italiana: due studi tassonomici a confronto. *Giornale Italiano di Psicologia*, 28:177–203.

[71] R. Dietz and A. Lang. Aefective agents: Effects of agent affect on arousal, attention, liking and learning. In *Cognitive Technology Conference*, 1999.

[72] Wen Dong, Bruno Lepri, and Alex (Sandy) Pentland. Modeling the coevolution of behaviors and social relationships using mobile phone data. In *Proc. of Mobile and Ubiquitous Multimedia*, MUM '11, pages 134–143, New York, NY, USA, 2011. ACM.

[73] M Brent Donnellan, Rand D Conger, and Chalandra M Bryant. The big five and enduring marriages. *Journal of Research in Personality*, 38(5):481–504, 2004.

[74] C. Dufour. *An investigation into the use of viral marketing for the companies and the key success factors of a good viral campaign*. PhD thesis, Dublin Business School, 2011.

[75] Nathan Eagle, Alex Pentland, and David Lazer. Inferring social network structure using mobile phone data. *Proceedings of the National Academy of Sciences (PNAS)*, 36(106):15274–15278, 2009.

[76] Nathan Eagle and Alex (Sandy) Pentland. Reality mining: sensing complex social systems. *Personal Ubiquitous Comput.*, 10(4):255–268, 2006.

[77] P. Ekman and W. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement.* Consulting Psychologists Press, 1978.

[78] P. Ekman and W. V. Friesen. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17:124–129, 1971.

[79] P. Ekman and W.V. Friesen. *Facial Action Coding System: Investigator's Guide*. Consulting Psychologists Press, Palo Alto, CA, 1978.

[80] Paul Ekman. *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage (Revised and Updated Edition)*. W. W. Norton & Company, 2001.

[81] A. Esuli and F. Sebastiani. SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC-2006*, pages 417–422, Genova, IT, 2006.

[82] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874, June 2008.

[83] T. Farma and I. Cortivonis. Un questionario sul ”locus of control”: suo utilizzo nel contesto italiano (a questionnaire on the locus of control: its use in the italian context. *Ricerca in Psicoterapia*, 2, 2000.

[84] W. Fleeson. Toward a Structure- and Process-Integrated View of Personality: Traits as Density Distributions of States. *Journal of Personality and Social Pyschology*, 80(6):1011–1027, 2001.

[85] William Fleeson. Situation-based contingencies underlying trait-content manifestation in behavior. *Journal of Personality*, 75(4):825–61, 2007.

[86] Joshua Fogel and Elham Nehmad. Internet social network communities: Risk taking, trust, and privacy concerns. *Computers in Human Behavior*, 25(1):153–160, 2009.

[87] B. J. Fogg and Dean Eckles. The behavior chain for online participation: how successful web services structure persuasion. In *Proceedings of the 2nd international conference on Persuasive technology*, PERSUASIVE'07, pages 199–209, 2007.

[88] James H. Fowler and Nicholas A. Christakis. Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the framingham heart study. *British Medical Journal*, 337, 2008.

[89] L. Freeman. Centrality in social networks: Conceptual clarification. *Social Networks*, 1(3):215–239, 1979.

[90] David C Funder. Towards a resolution of the personality triad: Persons, situations, and behaviors. *Journal of Research in Personality*, 40(1):21–34, 2006.

[91] D. Gatica-Perez. Automatic nonverbal analysis of social interaction in small groups: A review. *Image and Vision Computing*, 27(12), 2009.

[92] Peter A. Gloor, Kai Fischbach, Hauke Fuehres, Casper Lassenius, Tuomas Niinimki, Daniel Olguin Olguin, Sandy Pentland, Arttu Piri, and Johannes Putzke. Towards honest signals of creativity identifying personality characteristics through microscopic social network analysis. *Procedia - Social and Behavioral Sciences*, 26(0):166–179, 2011.

[93] Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. Predicting personality from twitter. In *SocialCom/PASSAT*, pages 149–156. IEEE, 2011.

[94] L R Goldberg. An alternative description of personality: the big-five factor structure. *Journal of Personality and Social Psychology*, 59(6):1216–1229, 1990.

[95] M C Gonzalez, C A Hidalgo, and A L Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):479–482, 2008.

[96] L.A. Goodman. Snowball sampling. *Annals of Mathematical Statistics*, 32:148–170, 1961.

[97] S Gosling. A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, 37(6):504–528, 2003.

[98] Lindsay T. Graham and Samuel D. Gosling. Can the ambiance of a place be determined by the user profiles of the people who visit it? In *ICWSM*, 2011.

[99] M.S. Granovetter. The Strength of Weak Ties. *The American Journal of Sociology*, 78(6):1360–1380, 1973.

[100] Marc Grootjen, Mark A. Neerincx, Jochum C. M. van Weert, and Khiet P. Truong. Measuring cognitive task load on a naval ship: implications of a real world environment. In *Proceedings of the 3rd international conference on Foundations of augmented cognition*, FAC'07, pages 147–156, Berlin, Heidelberg, 2007. Springer-Verlag.

[101] M. Guerini, A. Pepe, and B. Lepri. Do linguistic style and readability of scientific abstracts affect their virality. *Proceedings of ICWSM-12*, 2012.

[102] M. Guerini, O. Stock, and C. Strapparava. Valentino: A tool for valence shifting of natural language texts. In *Proceedings of LREC 2008*, Marrakech, Morocco, 2008.

[103] M. Guerini, C. Strapparava, and G. Özbal. Exploring text virality in social networks. In *Proceedings of ICWSM-11*, Barcelona, Spain, July 2011.

[104] Marco Guerini, Lorenzo Gatti, and Marco Turchi. Sentiment analysis: How to derive prior polarities from sentiwordnet. *Proceedings of EMNLP 2013*, pages 1259–1269, 2013.

[105] Marco Guerini, Jacopo Staiano, and Davide Albanese. Exploring image virality in google plus. In *Social Computing (SocialCom), 2013 International Conference on*, pages 671–678. IEEE, 2013.

[106] S Halko and J A Kientz. Personality and persuasive technology: An exploratory study on health-promoting mobile applications. pages 150–161. Persuasive, 2008.

[107] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18, 2009.

[108] Mark A Hall. *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato, 1999.

[109] Maureen T Hallinan and Warren N Kubitschek. The effects of individual and structural characteristics on intransitivity in social networks. *Social Psychology Quarterly*, 51(2):81–92, 1988.

[110] A. Hanjalic. Extracting moods from pictures and sounds: Towards truly personalized TV. *IEEE Signal Processing Magazine*, 2(23):90–100, 2006.

[111] A. Hanjalic, R. Lienhart, W. Y. Ma, and J. R. Smith. The holy grail of multimedia information retrieval: So close or yet so far away? *Proceedings of the IEEE*, 96(4):541–547, 2008.

[112] A. Hanjalic and Li-Qun Xu. Affective video content representation and modeling. *IEEE Transactions on Multimedia*, 7(1):143–154, 2005.

[113] Il-Horn Hann, Kai-Lung Hui, Sang-Yong Tom Lee, and Ivan PL Png. Overcoming online information privacy concerns: An information-processing theory approach. *Journal of Management Information Systems*, 24(2):13–42, 2007.

[114] Marc Hassenzahl and Noam Tractinsky. User experience - a research agenda. *Behaviour & Information Technology*, 25(2):91–97, 2006.

[115] E. Hatfield, J.T. Cacioppo, and R.L. Rapson. Emotional Contagion. *Current Directions in Psychological Science*, page 9699, 1993.

[116] R. Hazlett and J. Benedek. Measuring emotional valence to understand the user's experience of software. *International Journal of Human-Computer Studies*, 65(4):306–314, 2007.

[117] Alison L Hill, David G Rand, Martin A Nowak, and Nicholas A Christakis. Emotions as infectious diseases in a large social network: the SISa model. *Proceedings of the Royal Society B: Biological Sciences*, 277(1701):3827–3835, 2010.

[118] Jason I. Hong and James A. Landay. An architecture for privacy-sensitive ubiquitous computing. In *Proceedings of the 2Nd International Conference on Mobile Systems, Applications, and Services*, MobiSys '04, pages 177–189. ACM, 2004.

[119] Mariea Grubbs Hoy and George Milne. Gender differences in privacy-related measures for young adult facebook users. *Journal of Interactive Advertising*, 10(2):28–45, 2010.

[120] Bernardo A Huberman, Eytan Adar, and Leslie R Fine. Valuating privacy. *Security & Privacy, IEEE*, 3(5):22–25, 2005.

[121] Kai-Lung Hui, Hock Hai Teo, and Sang-Yong Tom Lee. The value of privacy assurance: An exploratory field experiment. *MIS Quarterly*, 31(1):19–33, 2007.

[122] D.P. Huijsmans and N. Sebe. How to complete performance graphs in content-based image retrieval: Add generality and normalize scope. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(2):245–251, 2005.

[123] H. Hung and D. Gatica Perez. Identifying dominant people in meetings from audio-visual sensors. In *IEEE Face and Gesture Recognition*, pages 1–6, 2008.

[124] Hayley Hung, Dinesh Babu Jayagopi, Sileye Ba, Jean-Marc Odobez, and Daniel Gatica-Perez. Investigating automatic dominance estimation in groups from visual attention and speaking activity. In *ICMI '08*, pages 233–236, 2008.

[125] T. I. Ianeva, A. P. de Vries, and H. Rohrig. Detecting cartoons: a case study in automatic video-genre classification. In *Proceedings of the 2003*

International Conference on Multimedia and Expo - Volume 2, ICME '03, pages 449–452, Washington, DC, USA, 2003. IEEE Computer Society.

[126] Remus Ilies, Michael D Johnson, Timothy A Judge, and Jessica Keeney. A within-individual study of interpersonal conflict as a work stressor: Dispositional and situational moderators. *Journal of Organizational Behavior*, 32(November 2009):44–64, 2011.

[127] Diana Zaiu Inkpen, Ol'ga Feiguina, and Graeme Hirst. Generating more-positive and more-negative text. In *Computing Attitude and Affect in Text: Theory and Applications*, pages 187–198. Springer, 2006.

[128] Katherine Isbister, Kia Höök, Jarmo Laaksolahti, and Michael Sharp. The sensual evaluation instrument: Developing a trans-cultural self-report measure of affect. *International Journal of Human-Computer Studies*, 65(4):315–328, 2007.

[129] A. Isen and A. Labroo. *Some Ways in Which Positive Affect Facilitates Decision Making and Judgment*, pages 365–93. 2003.

[130] A. Jaimes, D. Gatica-Perez, N. Sebe, and T.S. Huang. Human-centered computing: Towards a human revolution. *IEEE Computer*, 5(40):30–34, 2007.

[131] S. Jamali. Comment mining, popularity prediction, and social network analysis. Master's thesis, George Mason University, Fairfax, VA, 2009.

[132] Salman Jamali and Huzefa Rangwala. Digging digg : Comment mining, popularity prediction, and social network analysis. In *Proceedings of International Conference on Web Information Systems and Mining*, 2009.

[133] Carlos Jensen, Colin Potts, and Christian Jensen. Privacy practices of internet users: Self-reports versus observed behavior. *International Journal of Human-Computer Studies*, 63(1-2):203–227, 2005.

[134] Jing Jiang, Christo Wilson, Xiao Wang, Peng Huang, Wenpeng Sha, Yafei Dai, and Ben Y Zhao. Understanding latent interactions in online social networks. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, pages 369–382. ACM, 2010.

[135] O.P. John and S. Srivastava. *The Big Five trait taxonomy: History, measurement, and theoretical perspectives*, pages 102–138. 1999.

[136] Hideo Joho, Joemon M. Jose, Roberto Valenti, and Nicu Sebe. Exploiting facial expressions for affective video summarisation. In *ACM International Conference on Image and Video Retrieval (CIVR)*, 2009.

[137] Hideo Joho, Jacopo Staiano, Nicu Sebe, and Joemon M. Jose. Looking at the viewer: analysing facial activity to detect personal highlights of multimedia contents. *Multimedia Tools Appl.*, 51(2):505–523, 2011.

[138] Timothy A Judge, Daniel Heller, and Michael K Mount. Five-factor model of personality and job satisfaction: a meta-analysis. *Journal of Applied Psychology*, 87(3):530–541, 2002.

[139] Timothy A Judge and Remus Ilies. Affect and job satisfaction: a study of their relationship at work and at home. *Journal of Applied Psychology*, 89(4):661–673, 2004.

[140] Iris A Junglas, Norman A Johnson, and Christiane Spitzmüller. Personality traits and concern for privacy: an empirical study in the context of location-based services. *European Journal of Information Systems*, 17(4):387–402, 2008.

[141] D. Kahneman, A.B. Krueger, D.A. Schkade, N. Schwarz, and A.A. Stone. A survey method for characterizing daily life experience: The day reconstruction method. *Science*, 306:1776–1780, 2004.

[142] Daniel Kahneman, Barbara L. Fredrickson, Charles A. Schreiber, and Donald A. Redelmeier. When more pain is preferred to less: Adding a better end. *Psychological Science*, 4(6):401–405, 1993.

[143] Y. Kalish and G.L. Robins. Psychological predispositions and network structure: The relationship between individual predispositions, structural holes and network closure. *Social Networks*, 28(1):56–84, 2006.

[144] Alaina Kanfer and J.S. Tanaka. Unraveling the web of personality judgments: The influence of social networks on personality assessment. *Journal of Personality*, 61(4):711–738, 1993.

[145] H.B. Kang. Analysis of scene context related with emotional events. In *ACM International Conference on Multimedia*, 2002.

[146] D Keltner. Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame. journal of personality. *Social Psychology*, 68(3):441–454, 1995.

[147] D.A. Kenny. *Interpersonal perception: A social relations analysis*. The Guilford Press, 1994.

[148] Elham Khabiri, Chiao-Fang Hsu, and James Caverlee. Analyzing and predicting community preference of socially generated metadata: A case study on comments in the digg community. In *ICWSM*, 2009.

[149] Peter D Killworth and H Russell Bernard. The reversal small-world experiment. *Social networks*, 1(2):159–192, 1979.

[150] K.J. Klein, B.C. Lim, Saltz J.L., and D.M. Mayer. How do they get there? an examination of the antecedents of network centrality in team networks. *Academy of Management Journal*, 4:952–963, 2004.

[151] C.L. Kleinke. Gaze and eye contact: A research review. *Psychological Review*, 100:78–100, 1986.

[152] Bart P Knijnenburg, Alfred Kobsa, and Hongxia Jin. Dimensionality of information disclosure behavior. *International Journal of Human-Computer Studies*, 71(12):1144–1162, 2013.

[153] Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2*, IJCAI'95, pages 1137–1143, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

[154] Meera Komarraju and Steven J Karau. The relationship between the big five personality traits and academic motivation. *Personality and individual differences*, 39(3):557–567, 2005.

[155] Melinda Korzaan, Nita Brooks, and Timothy Greer. Demystifying personality and privacy: An empirical investigation into antecedents of concerns for information privacy. *Journal of Behavioral Studies in Business*, 1, 2009.

[156] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.

[157] N D Lane, E Miluzzo, H Lu, D Peebles, T Choudhury, and A T Campbell. A survey of mobile phone sensing. *IEEE Communications Magazine*, 2010.

[158] S.R.H. Langton. The mutual influence of gaze and head orientation in the analysis of social attention direction. *Q.J. Experimental Psychology*, 53A(3):825–845, 2000.

[159] S.R.H. Langton, R.J. Watt, and V. Bruce. Do the eyes have it? Cues to the direction of social attention. *Trends in Cognitive Sciences*, 4:50–59, 2000.

[160] V Latora and Marchiori M Phys Rev Lett. Efficient behavior of small-world networks. 2001.

[161] V Latora and M Marchiori. Economic small-world behavior in weighted networks. *Eur. Phys. Journ. B Condensed Matter*, 32(2), 2003.

[162] V Latora and M Marchiori. A measure of centrality based on the network efficiency. *New J. Phys. 9*, 188, 2007.

[163] David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-lászló Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. Computational Social Science. *Science*, 323(2):721–723, 2009.

[164] B. Lepri, N. Mana, A. Cappelletti, F. Pianesi, and M. Zancanaro. Modelling personality of participants during group interactions. In *UMAP*, 2009.

[165] Bruno Lepri, Jacopo Staiano, Giulio Rigato, Kyriaki Kalimeri, Ailbhe Finnerty, Fabio Pianesi, Nicu Sebe, and Alex Pentland. The sociometric badges corpus: A multilevel behavioral dataset for social behavior in complex organizations. In *SocialCom/PASSAT*, pages 623–628, 2012.

[166] Bruno Lepri, Ramanathan Subramanian, Kyriaki Kalimeri, Jacopo Staiano, Fabio Pianesi, and Nicu Sebe. Employing social gaze and speaking activity for automatic determination of the extraversion trait. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, page 7. ACM, 2010.

[167] Bruno Lepri, Ramanathan Subramanian, Kyriaki Kalimeri, Jacopo Staiano, Fabio Pianesi, and Nicu Sebe. Connecting meeting behavior with extraversion - a systematic study. *Affective Computing, IEEE Transactions on*, 3(4):443–455, 2012.

[168] Kristina Lerman and Aram Galstyan. Analysis of social voting patterns on digg. In *Proceedings of the first workshop on Online social networks*, WOSP '08, pages 7–12, New York, NY, USA, 2008. ACM.

[169] Kristina Lerman and Rumi Ghosh. Information contagion: an empirical study of the spread of news on digg and twitter social networks. In *Proceedings of 4th International Conference on Weblogs and Social Media (ICWSM-10)*, Mar 2010.

[170] Jurij Leskovec. *Dynamics of large networks*. ProQuest, 2008.

[171] Kurt Lewin, Fritz Trans Heider, and Grace M Heider. Principles of topological psychology. 1936.

[172] Hairong Li and Janice L Bukovac. Cognitive impact of banner ad characteristics: An experimental study. *Journalism & Mass Communication Quarterly*, 76(2):341–353, 1999.

[173] Jialiu Lin, Guang Xiang, Jason I. Hong, and Norman Sadeh. Modeling people's place naming preferences in location sharing. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*, pages 75–84. ACM, 2010.

[174] Janne Lindqvist, Justin Cranshaw, Jason Wiese, Jason Hong, and John Zimmerman. I'm the mayor of my house: examining why people use foursquare-a social-driven location sharing application. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2409–2418. ACM, 2011.

[175] B. Liu and L. Zhang. A survey of opinion mining and sentiment analysis. *Mining Text Data*, pages 415–463, 2012.

[176] Janice Lo. Privacy concern, locus of control, and salience in a trust-risk model of information disclosure on social networking sites. In *AMCIS*, page 110, 2010.

[177] Irene Lopatovska and Ioannis Arapakis. Theories, methods and current research on emotions in library and information science, information retrieval and human-computer interaction. *Inf. Process. Manage.*, 47(4):575–592, 2011.

[178] Y E Lu, S Roberts, P Lio, R Dunbar, and J Crowcroft. Size matters: variation in personal network size, personality and effect on information transmission. ICCSE, 2009.

[179] Anmol Madan, Manuel Cebrian, David Lazer, and Alex Pentland. Social sensing for epidemiological behavior change. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, Ubicomp '10, pages 291–300, New York, NY, USA, 2010. ACM.

[180] D. Madigan and J. York. Bayesian graphical models for discrete data. *International Statistical Review*, 63:215–232, 1995.

[181] Sascha Mahlke, Michael Minge, and Manfred Thüring. Measuring multiple components of emotions in interactive contexts. In *CHI '06 extended abstracts on Human factors in computing systems*, CHI EA '06, pages 1061–1066, New York, NY, USA, 2006. ACM.

[182] François Mairesse and Marilyn Walker. Automatic recognition of personality in conversation. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 85–88. Association for Computational Linguistics, 2006.

[183] François Mairesse and Marilyn Walker. Words mark the nerds: Computational models of personality recognition through language. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, pages 543–548, 2006.

[184] François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *J. Artif. Intell. Res.(JAIR)*, 30:457–500, 2007.

[185] N. Mana, B. Lepri, P. Chippendale, A. Cappelletti, F. Pianesi, P. Svaizer, and M. Zancanaro. A multimodal annotated corpus of consensus decision making meetings. *Language Resources and Evaluation*, 41(3):409–429, 2007.

[186] Karl Marx. Critical marginal notes on the article the king of prussia and social reform. by a prussian. *On Revolution, The Karl Marx Library*, 1:20, 1994.

[187] R.C. Mayer and J.H. Davis. The effect of the performance appraisal system on trust for management: a field quasi-experiment. *Journal of Applied Psychology*, 84:123–136, 1999.

[188] R.P. Mc Afee and J. Mc Millan. Auctions and bidding. *Journal of Economic Literature*, 25:699–738, 1987.

[189] C McCarty and H D Green. Personality and personal networks. *Sunbelt XXV*, 2005.

[190] A Mehra, M Kilduff, and D J Brass. The social networks of high and low self-monitors: Implications for workplace performance. *Administrative Science Quarterly*, 46(1):121–146, 2001.

[191] A. Mehrabian. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14(4):261–292, 1996.

[192] G.S. Mesch and G. Beker. Are norms of disclosure of online and offline personal information associated with the disclosure of personal information online? *Human Communication Research*, 36:570–592, 2010.

[193] Gilad Mishne. Experiments with mood classification in blog posts. In *Proceedings of ACM SIGIR 2005 Workshop on Stylistic Analysis of Text for Information Access*, volume 19, 2005.

[194] S. Moncrieff, C. Dorai, and S. Venkatesh. Affect computing in film through sound energy dynamics. In *ACM International Conference on Multimedia*, 2001.

[195] A. Money and H. Agius. Video summarisation: A conceptual framework and survey of the state of the art. *Journal of Visual Communication and Image Representation*, 19(2):121–143, February 2008.

[196] Arthur Money and Harry Agius. Feasibility of personalized affective video summaries. In *Affect and Emotion in Human-Computer Interaction*. Springer, 2008.

[197] Colum Mooney, Micheál Scully, Gareth J. F. Jones, and Alan F. Smeaton. Investigating biometric response for information retrieval applications. In *European Conference on Information Retrieval*, pages 570–574, 2006.

[198] Sai T. Moturu, Inas Khayal, Nadav Aharony, Wei Pan, and Alex Pentland. Using social sensing to understand the links between sleep, mood, and sociability. In *SocialCom/PASSAT*, pages 208–214, 2011.

[199] Min Mun, Shuai Hao, Nilesh Mishra, Katie Shilton, Jeff Burke, Deborah Estrin, Mark Hansen, and Ramesh Govindan. Personal data vaults: A

locus of control for personal data streams. In *Proceedings of the 6th International Conference*, Co-NEXT '10, pages 1–12, 2010.

[200] A. Neviarouskaya, H. Prendinger, and M. Ishizuka. Affect analysis model: novel rule-based approach to affect sensing from text. *Natural Language Engineering*, 17(1):95, 2011.

[201] Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. Textual affect sensing for sociable and expressive online communication. In AnaC.R. Paiva, Rui Prada, and RosalindW. Picard, editors, *Affective Computing and Intelligent Interaction*, volume 4738 of *Lecture Notes in Computer Science*, pages 218–229. Springer Berlin Heidelberg, 2007.

[202] Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. Compositionality principle in recognition of fine-grained emotions from text. In *ICWSM*, 2009.

[203] Martin A Nowak. Five rules for the evolution of cooperation. *science*, 314(5805):1560–1563, 2006.

[204] Jon Oberlander and Scott Nowson. Whose thumb is it anyway?: classifying author personality from weblog text. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 627–634. Association for Computational Linguistics, 2006.

[205] Daniel Olguın Olguın, Peter A Gloor, and Alex Sandy Pentland. Capturing individual and group behavior with wearable sensors. *Proceedings of the 2009 aaai spring symposium on human behavior modeling, SSS*, 9, 2009.

[206] D.O. Olguin, P.A. Gloor, and A. Pentland. Wearable sensors for pervasive healthcare management. In *Pervasive Computing Technologies for*

Healthcare, 2009. PervasiveHealth 2009. 3rd International Conference on, pages 1 –4, april 2009.

[207] Daniel Olguin-Olguin, Benjamin N. Waber, Taemie Kim, Akshay Mohan, Koji Ara, and Alex Pentland. Sensible organizations: technology and methodology for automatically measuring organizational behavior. *Trans. Sys. Man Cyber. Part B*, 39(1):43–55, February 2009.

[208] R Oliveira, A Karatzoglou, P Concejero, A Armenta, and N Oliver. Towards a psychographic user model from mobile phone usage. *ACM CHI, WIP*, pages 2191–2196, 2011.

[209] Elinor Ostrom. *Governing the commons: The evolution of institutions for collective action*. Cambridge university press, 1990.

[210] Paul Over, Alan F. Smeaton, and Philip Kelly. The TRECVID 2007 BBC rushes summarization evaluation pilot. In *TVS '07: Int. workshop on TRECVID video summarization*, pages 1–15, 2007.

[211] G. Ozbal and C. Strapparava. A computational approach to the automation of creative naming. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2012.

[212] G. Ozbal, C. Strapparava, and M. Guerini. Brand pitt: A corpus to explore the art of naming. In *Proceedings of LREC-2012*, 2012.

[213] G. Paltoglou, M. Thelwall, and K. Buckley. Online textual communications annotated with grades of emotion strength. In *Proceedings of the 3rd International Workshop of Emotion: Corpora for research on Emotion and Affect*, pages 25–31, 2010.

[214] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

[215] Maja Pantic and Leon J. M. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(12):1424–1445, 2000.

[216] A. Pentland. Social aware computation and communication. *IEEE Computer*, 38(3):33–40, 2005.

[217] Alex Pentland. Society's nervous system: Building effective government, energy, and public health systems. *IEEE Computer*, 45(1):31–38, 2012.

[218] Alex Pentland. *Social Physics: How Good Ideas SpreadThe Lessons from a New Science*. Penguin Press HC, The, January 2014.

[219] Alex Sandy Pentland. *Honest signals*. MIT press, 2010.

[220] D.I. Perrett and N.J. Emery. Understanding the intentions of others from visual signals: neurophysiological evidence. *Cashiers de Psychologie Cognitive*, 13:683–694, 1994.

[221] M Perugini and L Di Blas. *The Big Five Marker Scales (BFMS) and the Italian AB5C taxonomy: Analyses from an emic-etic perspective*. Hogrefe & Huber Publishers, 2002.

[222] Fabio Pianesi, Nadia Mana, Alessandro Cappelletti, Bruno Lepri, and Massimo Zancanaro. Multimodal recognition of personality traits in social interactions. In *Proceedings of the 10th international conference on Multimodal interfaces*, pages 53–60. ACM, 2008.

[223] Rosalind W Picard. Affective computing for hci. In *HCI (1)*, pages 829–833, 1999.

[224] I. Piller. 10. advertising as a site of language contact. *Annual Review of Applied Linguistics*, 23:170–183, 2003.

[225] Nataša Pržulj. Protein-protein interactions: Making sense of networks via graph-theoretic modeling. *Bioessays*, 33(2):115–123, 2011.

[226] Robert D Putnam. Bowling alone: America's declining social capital. *Journal of democracy*, 6(1):65–78, 1995.

[227] D. Quercia, J. Ellis, L. Capra, and J. Crowcroft. In the mood for being influential on twitter. *Proceedings of IEEE SocialCom'11*, 2011.

[228] D Quercia, M Kosinski, D Stillwell, and J Crowcroft. Our twitter profiles, our selves: Predicting personality with twitter. SocialCom/PASSAT, 2011.

[229] D Quercia, R Lambiotte, D Stillwell, M Kosinski, and J Crowcroft. The personality of popular facebook users. CSCW, 2012.

[230] Mika Raento, Antti Oulasvirta, and Nathan Eagle. Smartphones an emerging tool for social scientists. *Sociological methods & research*, 37(3):426–454, 2009.

[231] Javier Ramirez, Jose C. Segura, Carmen Benitez, Angel de la Torre, and Antonio Rubio. Efficient voice activity detection algorithms using long-term speech information. *Speech Communication*, 42(3-4):271 – 287, 2004.

[232] Javier Ramırez, José C Segura, Carmen Benıtez, Angel De La Torre, and Antonio Rubio. Efficient voice activity detection algorithms using long-term speech information. *Speech communication*, 42(3):271–287, 2004.

[233] David G Rand, Anna Dreber, Tore Ellingsen, Drew Fudenberg, and Martin A Nowak. Positive interactions promote public cooperation. *Science*, 325(5945):1272–1275, 2009.

[234] Niklas Ravaja, Marko Turpeinen, Timo Saari, Sampsa Puttonen, and Liisa Keltikangas-Järvinen. The psychophysiology of james bond: Phasic emotional responses to violent video game events. *Emotion*, 8(1):114, 2008.

[235] Byron Reeves and Clifford Nass. *How people treat computers, television, and new media like real people and places*. CSLI Publications and Cambridge university press, 1996.

[236] David N Reshef, Yakir A Reshef, Hilary K Finucane, Sharon R Grossman, Gilean McVean, Peter J Turnbaugh, Eric S Lander, Michael Mitzenmacher, and Pardis C Sabeti. Detecting novel associations in large data sets. *science*, 334(6062):1518–1524, 2011.

[237] Christopher Riederer, Vijay Erramilli, Augustin Chaintreau, Balachander Krishnamurthy, and Pablo Rodriguez. For sale : Your data: By : You. In *Proceedings of the 10th ACM Workshop on Hot Topics in Networks*, HotNets-X, pages 13:1–13:6, New York, NY, USA, 2011. ACM.

[238] S G B Roberts, R Wilson, P Fedurek, and R I M Dunbar. Individual differences and personal social network size and structure. *Personality and Individual Differences*, 44(4):954–964, 2008.

[239] Julian B Rotter. A new scale for the measurement of interpersonal trust. *Journal of personality*, 35(4):651–665, 1967.

[240] Andrew Ryan, Jeffery F Cohn, Simon Lucey, Jason Saragih, Patrick Lucey, Fernando De la Torre, and Adam Rossi. Automated facial expression recognition system. In *Security Technology, 2009. 43rd Annual 2009 International Carnahan Conference on*, pages 172–177. IEEE, 2009.

[241] P. Salovery, J. Mayer, and D. Rosenhan. Mood and helping: Mood as a motivator of helping and helping as regulator of moodl. *Review of Personality and Social Psychology*, 12:215–237, 1991.

[242] Jocelyn Scheirer, Raul Fernandez, Jonathan Klein, and Rosalind W Picard. Frustrating the user on purpose: a step toward building an affective computer. *Interacting with computers*, 14(2):93–118, 2002.

[243] Thomas C Schelling. *Micromotives and macrobehavior*. WW Norton & Company, 2006.

[244] Doris Schiöberg, Stefan Schmid, Fabian Schneider, Steve Uhlig, Harald Schiöberg, and Anja Feldmann. Tracing the birth of an osn: social graph and profile analysis in google+. In *Proceedings of the 3rd Annual ACM Web Science Conference*, WebSci '12, pages 265–274, New York, NY, USA, 2012. ACM.

[245] Johann Schrammel, Christina Köffel, and Manfred Tscheligi. Personality traits, usage patterns and information disclosure in online communities. In *Proceedings of the 23rd British HCI Group Annual Conference on People and Computers: Celebrating People and Technology*, pages 169–174. British Computer Society, 2009.

[246] Doc Searls. *The Intention Economy: When Customers Take Charge*. Harvard Business Press, 2013.

[247] N. Sebe, I. Cohen, F.G. Cozman, and T.S. Huang. Learning probabilistic classifiers for human-computer interaction applications. *Multimedia Systems*, 10(6):484–498, 2005.

[248] Jon F Sigurdsson. Computer experience, attitudes toward computers and personality characteristics in psychology undergraduates. *Personality and Individual Differences*, 12(6):617–624, 1991.

[249] M Simmons, Lada A Adamic, and Eytan Adar. Memes online: Extracted, subtracted, injected, and recollected. *ICWSM 2011*, 2011.

[250] R. Snow, B. O'Connor, D. Jurafsky, and A.Y. Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 254–263. Association for Computational Linguistics, 2008.

[251] Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1631–1642, 2013.

[252] Mohammad Soleymani, Guillaume Chanel, Joep J.M. Kierkels, and Thierry Pun. Affective ranking of movie scenes using physiological signals and content analysis. In *ACM workshop on Multimedia semantics*, pages 32–39, 2008.

[253] Hossein Azari Soufiani and Edo Airoldi. Graphlet decomposition of a weighted network. *Journal of Machine Learning Research - Proceedings Track*, 22:54–63, 2012.

[254] Jacopo Staiano and Marco Guerini. Depeche mood: a lexicon for emotion analysis from crowd annotated news. In *In Proceedings of the ACL conference*, 2014.

[255] Jacopo Staiano, Bruno Lepri, Nadav Aharony, Fabio Pianesi, Nicu Sebe, and Alex Pentland. Friends don't lie: inferring personality traits from social network structure. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 321–330. ACM, 2012.

[256] Jacopo Staiano, Bruno Lepri, Ramanathan Subramanian, Nicu Sebe, and Fabio Pianesi. Automatic modeling of personality states in small group interactions. In *Proceedings of the 19th ACM international conference on Multimedia*, MM '11, pages 989–992, New York, NY, USA, 2011. ACM.

[257] Jacopo Staiano, María Menéndez, Alberto Battocchi, Antonella De Angeli, and Nicu Sebe. Ux_mate: from facial expressions to ux evaluation. In *Proceedings of the Designing Interactive Systems Conference*, pages 741–750. ACM, 2012.

[258] Jacopo Staiano, Nuria Oliver, Bruno Lepri, Rodrigo de Oliveira, Michele Caraviello, and Nicu Sebe. Money walks: a human-centric study on the economics of personal mobile information. In *Proceedings of the 2014 ACM Conference on Ubiquitous Computing*, 2014.

[259] B. Staw and Yochi Cohen-Charash. The dispositional approach to job satisfaction: more than a mirage, but not yet an oasis. *Journal of Organizational Behavior*, 26:59–78, 2005.

[260] Rainer Stiefelhagen, Jie Yang, and Alex Waibel. Modeling focus of attention for meeting indexing. In *ACM MM*, pages 3–10, 1999.

[261] Rainer Stiefelhagen, Jie Yang, and Alex Waibel. Modeling focus of attention for meeting indexing based on multiple cues. *IEEE Transactions on Neural Networks*, 13:928–938, 2002.

[262] P.J. Stone, D.C. Dunphy, and M.S. Smith. *The General Inquirer: A Computer Approach to Content Analysis.* MIT press, 1966.

[263] C. Strapparava and A. Valitutti. WordNet-Affect: an affective extension of WordNet. In *Proc. of 4^{th} International Conference on Language Re-*

sources and Evaluation (LREC 2004), pages 1083 – 1086, Lisbon, May 2004.

[264] Carlo Strapparava and Rada Mihalcea. Semeval-2007 task 14: Affective text. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 70–74. Association for Computational Linguistics, 2007.

[265] Carlo Strapparava and Rada Mihalcea. Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on Applied computing*, pages 1556–1560. ACM, 2008.

[266] Pero Subasic and Alison Huettner. Affect analysis of text using fuzzy semantic typing. *Fuzzy Systems, IEEE Transactions on*, 9(4):483–496, 2001.

[267] Ramanathan Subramanian, Jacopo Staiano, Kyriaki Kalimeri, Nicu Sebe, and Fabio Pianesi. Putting the pieces together: Multimodal analysis of social attention in meetings. In *Proceedings of the International Conference on Multimedia*, MM '10, pages 659–662, New York, NY, USA, 2010. ACM.

[268] Ramanathan Subramanian, Yan Yan, Jacopo Staiano, Oswald Lanz, and Nicu Sebe. On the relationship between head pose, social attention and personality prediction for unstructured and dynamic group interactions. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 3–10. ACM, 2013.

[269] J.W. Sung, T. Kanade, and D.J. Kim. Pose robust face tracking by combining active appearance models and cylinder head models. *International Journal of Computer Vision*, 80(2):260–274, 2008.

[270] R J Swickert, J B Hittner, J L Harris, and J A Herring. Relationships among internet use, personality, and social support. *Computers in Human Behavior*, 18(4):437–451, 2002.

[271] Colin Swindells, Karon E MacLean, Kellogg S Booth, and Michael Meitner. A case-study of affect measurement tools for physical user interface design. In *Proceedings of graphics interface 2006*, pages 243–250. Canadian Information Processing Society, 2006.

[272] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307, 2011.

[273] Karen P Tang, Jason I Hong, and Daniel P Siewiorek. Understanding how visual representations of location feeds affect end-user privacy concerns. In *Proceedings of the 13th international conference on Ubiquitous computing*, pages 207–216. ACM, 2011.

[274] H. Tao and T.S. Huang. Connected vibrations: A modal analysis approach to non-rigid motion tracking. In *IEEE Conference on Compter Vision and Pattern Recognition*, pages 735–740, 1998.

[275] A Tellegen. Brief Manual for the Differential Personality Questionnaire. *Unpublished manuscript University of Minnesota Minneapolis*, pages 1010–1031, 1982.

[276] Stefano Teso, Jacopo Staiano, Bruno Lepri, Andrea Passerini, and Fabio Pianesi. Ego-centric graphlets for personality and affective states recognition. In *Social Computing (SocialCom), 2013 International Conference on*, pages 874–877. IEEE, 2013.

[277] D. Tjondronegoro, Yi-Ping Chen, and B. Pham. Highlights for more complete sports video summarization. *IEEE Multimedia*, 11(4):22–37, 2004.

[278] Eran Toch and Inbal Levi. Locality and privacy in people-nearby applications. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 539–548. ACM, 2013.

[279] TrackSocial. Optimizing Facebook Engagement, whitepaper. 2012.

[280] J. Tsai, P.G. Kelley, L.F. Cranor, and N. Sadeh. Location sharing technologies: Privacy risks and controls. *I/S: A Journal of Law and Policy for the Information Society*, 6(2):119–151, 2010.

[281] Janice Y Tsai, Serge Egelman, Lorrie Cranor, and Alessandro Acquisti. The effect of online privacy information on purchasing behavior: An experimental study. *Information Systems Research*, 22(2):254–268, 2011.

[282] Thomas S Tullis. Screen design. *Handbook of human-computer interaction*, 2:503–532, 1988.

[283] Marco Turchi, Martin Atkinson, Alastair Wilcox, Brett Crawley, Stefano Bucci, Ralf Steinberger, and Erik Van der Goot. Onts: optima news translation system. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 25–30. Association for Computational Linguistics, 2012.

[284] J.R.R. Uijlings, A.W.M. Smeulders, and R.J.H. Scha. Real-time visual concept classification. *Multimedia, IEEE Transactions on*, 12(7):665 – 681, nov. 2010.

[285] R. Valenti, Z. Yucel, and T. Gevers. Robustifying eye center localization by head pose cues. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 612–618, 2009.

[286] Roberto Valenti, Nicu Sebe, and Theo Gevers. Combining head pose and eye location information for gaze estimation. *Image Processing, IEEE Transactions on*, 21(2):802–815, 2012.

[287] Roberto Valenti, Jacopo Staiano, Nicu Sebe, and Theo Gevers. Webcam-based visual gaze estimation. In *Image Analysis and Processing–ICIAP 2009*, pages 662–671. Springer, 2009.

[288] Max Van Kleek, Daniel A. Smith, Nigel Shadbolt, and m.c. schraefel. A decentralized architecture for consolidating personal information ecosystems: The webbox. In *PIM*, pages 177–189, 2012.

[289] Hans van Kuilenburg, Marco Wiering, and Marten den Uyl. A model based method for automatic facial expression recognition. In *Proceedings of the 16th European conference on Machine Learning*, ECML'05, pages 194–205, Berlin, Heidelberg, 2005. Springer-Verlag.

[290] C. J. van Rijsbergen. *Information Retrieval, Second Edition*. Butterworths, 1979.

[291] Hal Varian, Fredrik Wallenberg, and Glenn Woroch. The demographics of the do-not-call list. *IEEE Security & Privacy*, 3:34–39, 2005.

[292] Paul Viola and Michael J. Jones. Robust real-time face detection. *Int. J. Comput. Vision*, 57(2):137–154, May 2004.

[293] Jeffrey R Vittengl and Craig S Holt. A time-series diary study of mood and social interaction. *Motivation and Emotion*, 22(3):255–275, 1998.

[294] Michael Voit and Rainer Stiefelhagen. Deducing the visual focus of attention from head pose estimation in dynamic multi-view meeting scenarios. In *ICMI '08*, pages 173–180, 2008.

[295] H.L. Wang and L.F. Cheong. Affective understanding in film. *IEEE Transactions on Circuits and Systems for Video Technology*, 16(6):689–704, 2006.

[296] S. Wang and C.D. Manning. Baselines and bigrams: Simple, good sentiment and topic classification. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2012.

[297] Roy Want, Trevor Pering, Gunner Danneels, Muthu Kumar, Murali Sundar, and John Light. The personal server: Changing the way we think about ubiquitous computing. In *In Proceedings of 4th International Conference on Ubiquitous Computing*, pages 194–209, 2002.

[298] Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45(4):1191–1207, 2013.

[299] S Wasserman and K Faust. *Social Network Analysis: Methods and Applications*. Cambridge Press, 1994.

[300] David Watson, Lee A Clark, and Auke Tellegen. Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, 54(6):1063–1070, 1988.

[301] D J Watts and S Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.

[302] S Wehrli. Personality on social network sites: An application of the five factor model. *ETH Zurich Sociology Working Papers*, 7, 2008.

[303] Steve J Westerman, EJ Sutherland, L Robinson, H Powell, and G Tuck. A multi-method approach to the assessment of web page designs. In *Af-*

fective Computing and Intelligent Interaction, pages 302–313. Springer, 2007.

[304] L.R. Wheeless and J. Grotz. Conceptualization and measurement of reported self-disclosure. *Human Communication Research*, 2:338–346, 1976.

[305] Simon Whitehead and Lawrence Cavedon. Generating shifting sentiment for a conversational agent. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 89–97, Los Angeles, CA, June 2010. Association for Computational Linguistics.

[306] A. Whiten. Evolutionary and developmental origins of the mindreading system. In *Evolution and Development*. Lawrence Erlbaum, 1997.

[307] Steve Whittaker, David Frohlich, and Owen Daly-Jones. Informal workplace communication: what is it like and how might we support it? In *Conference on Human Factors in Computing Systems CHI*, volume Boston, US, pages 131–137. ACM, 1994.

[308] Jason Wiese, Patrick Gage Kelley, Lorrie Faith Cranor, Laura Dabbish, Jason I Hong, and John Zimmerman. Are you close with me? are you nearby?: investigating social groups, closeness, and willingness to share. In *Proceedings of the 13th international conference on Ubiquitous computing*, pages 197–206. ACM, 2011.

[309] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Vancouver, Canada, 2005.

[310] T. Wilson, J. Wiebe, and R. Hwa. Just how mad are you? finding strong and weak opinion clauses. In *Proceedings of AAAI*, pages 761–769, 2004.

[311] L. Wu, B.N. Waber, S. Aral, E. Brynjolfsson, and A. Pentland. Mining Face-to-Face Interaction Networks Using Sociometric Badges: Predicting Productivity in an IT Configuration Task. 2008.

[312] J. Xiao, T. Kanade, and J. Cohn. Robust full motion recovery of head by dynamic templates and re-registration techniques. In *IEEE Face and Gesture Recognition*, 2002.

[313] J. Xiao, T. Kanade, and J. Cohn. Accurate eye center location and tracking using isophote curvature. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2008.

[314] M. Xu, L.T. Chia, and J. Jin. Affective content analysis in comedy and horror videos by audio emotional event detection. In *IEEE International Conference on Multimedia and Expo*, 2005.

[315] Alyson L Young and Anabel Quan-Haase. Information revelation and internet privacy concerns on social network sites: a case study of facebook. In *Proceedings of the fourth international conference on Communities and technologies*, pages 265–274. ACM, 2009.

[316] Bieke Zaman and Tara Shrimpton-Smith. The facereader: Measuring instant fun of use. In *Proceedings of the 4th Nordic conference on Human-computer interaction: changing roles*, pages 457–460. ACM, 2006.

[317] Gloria Zen, Negar Rostamzadeh, Jacopo Staiano, Elisa Ricci, and Nicu Sebe. Enhanced semantic descriptors for functional scene categorization. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 1985–1988. IEEE, 2012.