

Preface

Online Social Media (OSM) have revolutionized the way people interact and share information. Many recent events and developments have shown that OSM are very powerful tools for people to organize and take action. Examples include the ‘Occupy’ movement, ‘Sandy’ relief efforts, and the ‘Arab Spring’. OSM have offered a real and viable alternative to conventional mainstream media. The latter is often accused of being biased. Spin and constraints imposed by regulating or funding bodies can hinder mainstream media outlets’ unbiased reporting. They often omit (intentionally or otherwise) certain details in their reporting. On the other hand, OSM are likely to provide “raw”, unedited information and the details can be overwhelming with the potential of misinformation and disinformation. Yet, OSM are leading to the democratization of knowledge and information. OSM is allowing almost any citizen to become a journalist reporting on specific events of interest. This is resulting in unimaginable amounts of information being shared among huge numbers of OSM participants. As of this writing, twitter claims to have 271 million monthly active users, producing 500 million tweets per day. Facebook grew by the end of 2013 to 1.23 billion users with 757 million users logging on every day. Facebook now has a user base comparable to the population of India! The Daily Mail further reports that the average American daily spends 40 min on Facebook. This is resulting in several billion “likes” and several 100 million posted pictures in a single day.

This book contains 10 contributions that tackle challenges of the subject of OSM analysis and visualization from different angles. These challenges include:

1. Many details are hidden in OSM posts and as a result, search engines tend to favor conventional media sources. Mechanisms that allow for delving deeper into the content of OSM posts so that a more meaningful search can be performed are needed. How can sources of a Twitter event, for example, be accurately identified?
2. Word-of-mouth marketing is a byproduct of OSM. However with OSM, there is the challenge of linking opinions of users to their demographic information. How can we infer such information with the little clues available on a social

network, such as Twitter? Can the gender of users be inferred based on colors used in Twitter profiles?

3. Interests of OSM participants change over time. How can we analyze this change and how can we present it in a convenient manner? What factors are essential for trend-prediction on Twitter, for instance?
4. With the large sizes of users personal networks (an average Facebook user has more than 220 friends), better ways for OSM users are needed in order to better view their social networks. How can the network be visualized in such a way that gives the user immediate important information about friends and posts, such as the level of activity of a friend and the popularity of a post?
5. Making privacy decisions (what to share with whom) is cumbersome given the sizes of personal networks. Going through the list of friends, one-by-one, may no longer be a viable solution. How can we make such privacy decisions aided by a visualization and categorization mechanism offering different privacy levels for different categories?
6. Network analysis is a complex process. This challenge is approached on two fronts: how can we (a) develop realistic models that describe people's interaction and (b) defuse the complexities of the processes through the development of appropriate tool?

Summary of Contributions

Identifying Event-Specific Sources from Social Media

Identifying valuable sources of social media events is a very challenging task, but is a necessary step toward identifying misinformation and disinformation in social media. These sources are often buried in the “long tail.” A quick search for some event on major search engines, such as Google, yields top hits for mainstream and conventional media. In addition, conventional media does not often include as much details as social media alternatives. In this chapter, Debanjan Mahata and Nitin Agarwal take on the challenging task of identifying sources from social media for specific events. The challenges include sparsity of resources, quality assessment, entity extraction, and evaluation measures.

Mahata and Agrawal develop a mutual reinforcement-based methodology for this identification, present an evaluation strategy, validate the approach using real-life data, and conclude with analysis of the model. The empirical evaluation is based on a data set of 11,378 blog posts from different blogging sources. The events are the Egyptian, Libyan, and Tunisian uprisings. Empirical results show that the developed evolutionary mutual reinforcement converges faster with better accuracy than the conventional mutual reinforcement model, and it outperforms Google Blog Search and IceRocket.

Demographic and Psychographic Estimation of Twitter Users Using Social Structures

With the ever expanding number of social media users, such as Twitter and Facebook, many of their posts are geared toward expressing opinions about certain products and services. This can provide a low-cost, real-time word-of-mouth marketing, as opposed to expensive, formal customer surveys. However, formal surveys have the advantage of linking opinions to customer attributes (such as age and gender), but such attributes are often hidden in social media. In this chapter, Jun Ito, Kyosuke Nishida, Takahide Hoshide, Hiroyuki Toda, and Tadasu Uchiyama analyze more than 4.6 million Twitter users in Japan. It is determined that very few users use values in their Twitter profiles that can reveal their age (roughly 3 % described their age), gender (less than 8 % revealed their gender), location (less than 25 % revealed their location), and occupation (less than 14 % indicated their job). Hence, the estimation of these attributes is necessary.

To address this limitation, Ito et al. provide in this chapter a method by which hidden attributes can be estimated from the publicly available information, Twitter profiles and posts, and from social neighbors. Specifically, Ito et al. estimate the four attributes: gender, age, occupation, and interest. The estimation proceeds at three levels. First, a labeling method that identifies users with blog accounts is used to extract their attributes from the blog profile as true labeling for the training data set. Experiments confirm that this is a more accurate labeling than other methods, such as manual labeling and pattern matching. Second, the authors investigate how to combine bag-of-words features of profile documents and tweets. Nine different combining methods are evaluated, identifying the best two of these methods. Finally, information from social neighbors is utilized. Three adjustment levels are investigated.

Say It With Colors: Language-Independent Gender Classification on Twitter

Gender prediction of social network users is important for targeted advertising, law enforcement, and other social reasons. Gender classification in networks such as Twitter heavily depends on analyzing the text of posted messages or tweets. Among the limitations of such an approach is language-dependence, intractability, and non-scalability.

In this chapter, Jalal Alowibdi, Ugo Buy, and Philip Yu present a gender classification approach that is based on colors. This approach is based on analyzing five color features used in Twitter user profiles: background, text, link, sidebar fill, and sidebar boarder colors. Colors are language independent, and using only five features in the analysis (as opposed to millions of features used in text-based classification) makes the color-based approach more desirable for scalability and

tractability. Realizing that the number of colors can be technically enormous, while practically there can be different grades of the same color, Alowibdi et al. employ a preprocessing step that converts the colors from their RGB (Red, Green, Blue) representation to HSV (Hue, Saturation, Value) representation. The colors are then sorted by their hue and value attributes, providing similar labellings for colors, which are then converted back to their RGB values. This preprocessing step helps improve accuracy and reduce the size of the data set. Empirical results show that the classification accuracy is roughly between 70 and 74 % for different data sets, which is a clear improvement over the 50 % norm. These results are obtained from a data set of about 170,000 users where it was possible to independently verify their genders.

TUCAN: Twitter User Centric Analyzer

This chapter by Luigi Grimaudo, Han Hee Son, Mario Baldi, Marco Mellia, and Maurizio Munafò takes a text-mining approach to analyzing Twitter posts, using a framework called TUCAN. TUCAN analyzes the tweets of a single, target user over a specific period of time, identifying the interests of that user during the given time window. TUCAN also offers the ability to do a comparison between several users, inferring any common interests. The results are depicted graphically in an intuitive visual representation.

The steps taken in this framework start by projecting a target user's tweets to a time window that Grimaudo et al. call a "bird song". Next, bird songs are filtered and cleaned using known methods in order to eliminate noise and derive general concept terms for the words in the songs. Terms are then scored to identify the important terms for a target user in a bird song. Finally, similarity scores are used to compare two bird songs. The results are provided visually, using colored matrices, where colors distinguish similarity scores.

The authors validate TUCAN by providing an empirical study that includes 740 Twitter users, including 28 public figures, over a period that exceeds two months. This results in analyzing more than 800,000 tweets. Grimaudo et al. also perform a parameter-sensitivity analysis of TUCAN.

Evaluating Important Factors and Effective Models for Twitter Trend Prediction

In this chapter, Peng Zhang, Xufei Wang, and Baoxin answer two important questions related to trend prediction in Twitter. The first question is what content and context factors (or a combination of them) are more important to Twitter trend prediction. The second question is which (if any) is more appropriate for prediction.

To answer these questions, Zhang et al. performed an empirical study using 16.8 million tweets by about 670,000 users over a period of several months. They conducted relevance analysis using tweet content, network topology, and user behavior, addressing the first question. To address the second question, they also performed a prediction performance study for several known prediction models. The analysis concluded that trend factors based on user behavior are more effective in predicting trends, and that the nonlinear state-space models are more suited for prediction.

Rings: A Visualization Mechanism to Enhance the User Awareness on Social Networks

The visualization of someone's social network activities on Facebook is the subject of this chapter by Shi Shi, Thomas Largillier, and Julita Vassileva. The authors take an approach to represent such activities using rings, where the colors and sizes of these rings represent different levels of activities. The result is a tool called *Rings*. *Rings* allows a user visually and interactively (1) see basic post information, such as the posters' identification and the time; (2) review the activity level of a user; and (3) assess the popularity of posts.

Shi et al. validate *Rings* using two user studies. The first study assesses if indeed *Rings* increases user awareness, whether it is useful to users, and how usable the user interface is. The empirical data show general positive results. The second user study delves deeper into *Rings* validating more specific properties of the system, such as performance, colors, reliability, and other factors.

Friends and Circle—A Design Study for Contact Management in Egocentric Online Social Networks

Due to the large amount of information that a social network user must deal with, managing privacy decisions related to which posts to share with which users becomes an overwhelming process. Users often indicate that they regret certain social network postings. In addition, an average user can easily have more than one hundred friends or followers; for example, a Facebook user had an average of 229 friends in 2013. Going through someone's network, friend-by-friend, in order to decide what level of privacy is required for each friend is not practical. Instead, users tend to categorize their networks allowing a different privacy-level for each category.

Bo Gao and Bettina Berendt target this issue in this chapter by developing an online application, called *FreeBu*, which visualizes a user's network and categorizes his/her friends. Gao and Berendt accomplish this task through several steps.

They present *CircleTree*, a visualization tool that incorporates modularity-based community detection (MOD). A user study is then performed to compare hierarchical MOD with Facebook smart lists. The findings show that hierarchical MOD provides more support for visibility decisions. They also show that graph-based algorithms for community detection are more appropriate than attribute-based algorithms. The authors, then, empirically compare MOD with Generative Model for Friendships (GMF). The study involves ego-networks as follows: 10 from Facebook, 909 from Twitter, and 129 from Google+. Using this data set, it is shown that MOD outperforms GMF. Finally, Gao and Berendt enrich *CircleTree* with three additional visual interactive methods culminating in *FreeBu*, exploiting their empirical findings.

Genetically Optimized Realistic Social Network Topology Inspired by Facebook

There is an ever-increasing need to better understand the topological properties of social networks. This requires the development of abstract and generic, and at the same time, flexible and realistic models that describe how people socially interconnect. The synthetic generation of such a topology is very handy to researchers since it provides them with a mechanism to generate social network data with certain specified properties on demand. In this chapter, Alexandru Topirceanu, Mihai Udrescu, and Mircea Vladutiu address this problem of synthetically generating realistic social network topologies. They propose the Genetic-Optimized Social Network (*Genosian*) method. The aim of *Genosian* is to create accurate replica of friendship models collected from Facebook.

The first empirical observation made by the authors is that realistic Facebook networks share common metrics, in spite of the fact that the networks are diverse in shape and size. It is found that topological metrics of these realistic Facebook networks fall within narrow thresholds. These metrics include: network size, average path length, clustering coefficient, average degree, diameter, density, and modularity. In addition, distribution of degrees, betweenness, closeness, and centrality are looked at.

Genosian uses a Genetic Algorithm (GA) to generate the social network. It starts with a random creation of a collection of communities, inspired by the Watts-Strogatz algorithm. Then, these communities are linked together. The GA optimizes these intra-community edges until the centrality measure of the graph is finessed, aiming at a comparable value to that of real-life Facebook examples. The rewiring of intra-community edges is carried out using GA's natural selection. The empirical results show that on average *Genosian* produces 63 % more accuracy than the best previous known method.

A Workbench for Visual Design of Executable and Re-usable Network Analysis Workflows

Network analysis is in general a complex process that consists of several steps. Providing social network researchers with tools to assist them in the analysis processes defuses some of these complexities. Tilman Göhnert, Andreas Harrer, Tobias Hecking, and H. Ulrich Hoppe provide in this chapter a social network analysis tool, called *Analytic Workbench*, which offers several advantages over similar tools. The motivation for the Analytic Workbench is rooted in ease of accessibility, support for complex analysis processes in an integrated environment, and explicit representation of analysis workflows. Another motivating factor is the support and ease of integration of additional analysis functions.

The result is a Web-based interface that provides visual representation of multi-step analysis workflows. Explicit representation of analysis workflows yields the ability to reuse workflows, allowing comparative studies with the same analytic methodology. Furthermore, the tool easily provides adaptation of parts of a workflow in another, allowing a researcher to experiment with different algorithms and analytic steps.

Göhnert et al. showcase the Analytic workbench by using blockmodeling analysis on multi-relational networks. The authors also report on a user evaluation study of the tool.

On the Usage of Network Visualization for Multiagent System Verification

In this chapter, Fatemeh H. Fard and Behrouz H. Far take advantage of visualization techniques in order to build an approach for the verification of Multi-Agent Systems (MAS). They make use of social network analysis in order to model the interaction of agents. The purpose of this study is to detect emergent behavior in these networks. An emergent event is an unexpected run-time behavior of the system unforeseen by system designers.

Fard and Far avoid using model checking techniques for detecting emergent behavior due to the scalability drawback of these approaches. Instead with the use of interaction matrices, three networks are derived. The first is a component-level network that describes the agent's behavior. The other two are system-level networks, defining the interaction of agents. The authors illustrate this approach and compare it to other existing approaches through two examples.

Online Social Media Analysis and Visualization

Kawash, J. (Ed.)

2014, XVI, 233 p. 94 illus., 76 illus. in color., Hardcover

ISBN: 978-3-319-13589-2