Data-Driven Modeling and Analysis of Online Social Networks

Divyakant Agrawal[1], Bassam Bamieh[2], Ceren Budak[1], Amr El Abbadi[1],
Andrew Flanagin[3], and Stacy Patterson[2]

[1] Department of Computer Science
University of California at Santa Barbara
Santa Barbara, CA 93106, USA
[2] Department of Mechanical Engineering
University of California at Santa Barbara
Santa Barbara, CA 93106, USA
[3] Department of Communication
University of California at Santa Barbara
Santa Barbara, CA 93106, USA

Abstract. With hundreds of millions of users worldwide, social networks provide incredible opportunities for social connection, learning, political and social change, and individual entertainment and enhancement in a wide variety of forms. In light of these notable outcomes, understanding information diffusion over online social networks is a critical research goal. Because many social interactions currently take place in online networks, we now have have access to unprecedented amounts of information about social interaction. Prior to the advent of such online networks, investigations about social behavior required resource-intensive activities such as random trials, surveys, and manual data collection to gather even small data sets. Now, massive amounts of information about social networks and social interactions are recorded. This wealth of data can allow us to study social interactions on a scale and at a level of detail that has never before been possible.

We present an integrated approach to information diffusion in online social networks focusing on three key problems: (1) Querying and analysis of online social network datasets; (2) Modeling and analysis of social networks; and (3) Analysis of social media and social interactions in the contemporary media environment. The overarching goals are to generate a greater understanding of social interactions in online networks through data analysis, to develop reliable and scalable models that can predict outcomes of these social processes, and ultimately to create applications that can shape the outcome of these processes. We start by developing and refining models of information diffusion based on real-world data sets. We next address the problem of finding influential users in this data-driven framework. It is equally important to identify techniques that can slow or prevent the spread of misinformation, and hence algorithms are explored to address this question. A third interest is the process by which a social group forms opinions about an idea or product, and we therefore describe preliminary approaches to create models that

H. Wang et al. (Eds.): WAIM 2011, LNCS 6897, pp. 3–17, 2011.

accurately capture the opinion formation process in online social networks. While questions relating to the propagation of a single news item or idea are important, these *information campaigns* do not exist in isolation. Therefore, our proposed approach also addresses the interplay of the many information diffusion processes that take place simultaneously in a network and the relative importance of different topics or *trends* over multiple spatial and temporal resolutions.

Keywords: Information propagation, Social Networks, Data Analysis, Sub-modular optimization.

1 Introduction

Internet technologies and online social networks are changing the nature of social interactions. There exist incredible opportunities for learning, social connection, and individual entertainment and enhancement in a wide variety of forms. People now have ready access to almost inconceivably vast information repositories that are increasingly portable, accessible, and interactive in both delivery and formation. Basic human activities have changed as a result, and new possibilities have emerged. For instance, the process by which people locate, organize, and coordinate groups of individuals with shared interests, the number and nature of information and news sources available, and the ability to solicit and share opinions and ideas across myriad topics have all undergone dramatic change as a result of interconnected digital media. Indeed, recent evidence indicates that 45% of users in the U.S. say that the Internet played a crucial or important role in at least one major decision in their lives in the last two years, such as attaining additional career training, helping themselves or someone else with a major illness or medical condition, or making a major investment or financial decision [HR06]. The significant role of online social networks in human interactions motivates our goals of generating a greater understanding of social interactions in online networks through data analysis, the development of reliable models that can predict outcomes of social processes, and ultimately the creation of applications that can shape the outcome of these processes.

This work centers on social processes related to the diffusion of information and opinions in online social networks. We are interested in questions relating to how a single news item or idea propagates through a social network. We also note that such “information campaigns” do not exist in isolation. So while it is important to understand the dynamics of individual campaigns, we also study the interplay of the many information diffusion processes that take place simultaneously in a network and the relative importance of different information topics or *trends* over multiple geographical and temporal resolutions.

Diffusion of Information and Opinions. Given the notable outcomes and the potential applications of information diffusion over online social networks, it is an important research goal to increase our understanding of information diffusion processes. It is also desirable to understand how these processes can be modified to achieve desired objectives. We describe and refine models of information diffusion for online social networks based on analysis of real-world data sets. The scalable, flexible identification of influential users or opinion leaders for specific topics is crucial to ensuring that information reliably reaches a large audience, and so we propose to develop algorithms for finding influential users in this data-driven framework. It is equally important to identify techniques that can slow or prevent the spread of misinformation, and we create models and algorithms to address this question. We are also interested in the processes by which a social group forms opinions about an idea or product. We create models that accurately capture the opinion formation process in online social networks and develop scalable algorithms and techniques for external intervention that can alter those opinions.

Information Trend Analysis. In light of an online environment where there is an increasing number of sources of information, understanding information trends is critical. For example, data showing that there is sustained and substantial interest in a particular political issue locally, but not regionally, nationally, or internationally, can signal opportunities for political organization around that topic, political party proclivities in a geographic area, or even where best to focus fund-raising efforts. The relative trendiness of a message or topic, such as when interest in it suddenly spikes, can be more important than the overall popularity of a topic. Finally, underlying trends that illustrate relationships among entities can be used to highlight information that is otherwise obscure. For example, strangers independently confirming the importance of a certain piece of information (such as a user's feedback rating on eBay) are better indicators of its veracity than are a tightly-connected cluster of information sources reporting the same information in common. Trending tools can thus help to discern these and other distinctions. We are interested in developing tools to discover many types of trends at multiple spatio-temporal resolutions.

In the following sections, we detail our research questions, propose solution techniques, and highlight open problems for further investigation.

2 Information Diffusion

Social networks have already emerged as a significant medium for the widespread distribution of news and instructions in mass convergence events such as the 2008 U.S. Presidential Election [HP09], the 2009 Presidential election in Iran [Gro09], and emergencies like the landfall of Hurricanes Ike and Gustav in the fall of 2008 [HP09]. Due to the open nature of social networks, it is not only possible to spread factual information, but also to create misinformation campaigns that can discourage or disrupt correct response in these situations. For instance, Twitter has served as forum for spreading both useful information and false rumors about the H1N1 pandemic which altered users' attitudes about the safety and efficacy of the flu vaccine [Far09].

Given the notable outcomes and the potential applications of information diffusion over online social networks, it is an important research goal to increase our understanding of information diffusion processes and to study how these processes can be modified to achieve desired objectives. We have developed preliminary models of information and opinion diffusion for online social networks. First, we address the question of how to formalize a loosely specified model of social diffusion based on data analysis in order to adapt this model to social interactions in the blogosphere. We then consider the problem of limiting the reach of an information campaign and highlight preliminary results and proposed extensions.

2.1 Influence Maximization: The Law of the Few

The identification of *influential users* or *opinion leaders* in a social network is a problem that has received a significant amount of attention in recent research [KKT03, CYY09, LKG$^+$07, KSNM09, WCSX10]. Informally, we define an *information campaign* as the process by which a news item or idea spreads from the initiators of the campaign, *i.e.* the users who first learn the news, throughout the social network. This initial set of users is denoted by the set A, and the set of users who eventually receive this news is the *influence set* of A, denoted $IF(A)$. In the *influence maximization problem*, given a model of how information diffuses in a social network, the objective is to select a set of users A of size k who are to be the initial recipients of the information (through some offline means), so that the size of $IF(A)$ is maximized [DR01, RD02].

With an efficient, robust solution to this problem, it would be possible to ensure the widespread dissemination of important information in a social network. Early works relied on heuristics such as node degree and distance centrality [WF94] to select the set A. More recently, several investigators have considered probabilistic models of information diffusion such as the *Independent Cascade* [GLM01] and *Linear Threshold* [Gra78]. The problem of finding the optimal initiator set in this model is NP-hard, but there is a polynomial-time greedy algorithm that yields a result that is within $1-1/e$ of optimal [KKT03]. Work has been done on improving the performance of greedy algorithms for influence maximization [CYY09, LKG$^+$07, KSNM09, WCSX10], but the scalability of these algorithms, in the context of social networks spanning millions of users, remains a significant challenge.

We summarize our recent work [BAEA10] on the influential users problem using a different model of diffusion based on the theories introduced in the popular book "The Tipping Point" by Malcolm Gladwell [Gla02]. The main idea of "The Tipping Point" is the crucial roles of three types of "fascinating" people that the author calls *mavens, connectors* and *salesmen* on the effectiveness of an information campaign. These people are claimed to "play a critical role in the word-of-mouth epidemics that dictate our tastes, trends and fashions". We study those three types of people or *actors* and a fourth type of interesting actor that we call a *translator.*

The first type of actor introduced by Gladwell is the connector. In terms of a social network graph, we define a connector to be a node that has high degree centrality. The second type of important actor in information propagation is the maven. The word "maven" comes from Yiddish and means one who accumulates knowledge. Gladwell lists three important characteristics for mavens: 1) they seek new knowledge, 2) they share the knowledge they acquire with others and 3) an individual hearing something from a maven is very likely to believe the correctness and importance of this piece of information. Translating those features into graph theory, we define mavens to be nodes that start a large number of cascades (they are the original source of new information) and have high influence on their neighbors. The third kind of actor that Gladwell introduces is the salesman, a person with high charisma who can sell ideas to almost anyone since he never gives up. We define a salesman to be a node that has a large number of trials to activate its neighbors for cascades that the node itself is a part of. We also study another class of actors referred to as *translators*. These actors act as bridges or translators among different communities and therefore have the power of changing the context in which to present an idea. In order to identify the translators in the blogosphere, first we need to detect the communities. Different from many community detection research [Cla05, PSL90, New03, New06, GL08, BGMI05, ZWZ07, Gre07], we define communities based on the existence of flow of influence between nodes rather than relying solely on the structure of the graph. Using this definition, we detect overlapping communities using the algorithm presented in [BGMI05] and define translators as the nodes that belong to the most number of communities.

Weblogs have become a predominant way of sharing data online. The blogosphere has considerable influence on how people make decisions in areas such as politics or technology [AG05]. We used the August-October 2008 Memetracker data that contains timestamped phrase and link information for news media articles and blog posts from different blogs and news websites. The data set consists of 53,744,349 posts and 2,115,449 sources of information (blogs, news media and sources that reside outside the blogosphere) and 744,189 cascades. Using the methods formalized above, we identify the mavens, salesmen, connectors and translators of the blogosphere and study their effect on the success of cascades. Our initial results are quite promising in that they indicate that algorithms for finding the best method of reaching out to certain actors, rather than the entire network, can be a good heuristic to impact influence in social networks. The types of actors identified can also be used to augment the current models of diffusion to capture real world behavior. As part of future work, we also plan to augment our analysis on the intermediaries to investigate if there exists an optimal timing to reach out to a connector, maven, salesman or translator. Are these actors more useful if they adopt and advocate a cascade early or later on? We analyzed the blogosphere data to investigate the validity of the heuristics introduced but the same heuristics can indeed be evaluated on other social networks. We believe that different social networks provide different types of interactions, which means that certain actors, while not so significant in some networks, can be highly influential in others.

2.2 Limiting Diffusion of Misinformation

While a substantial amount of research has been done in the context of influence maximization, a problem that has not received much attention is that of limiting the influence of a malicious or incorrect information campaign. One strategy to deal with a misinformation campaign is to limit the number of users who are willing to accept and spread this misinformation. In this section, we present preliminary work on techniques to limit the number of users who participate in an information campaign. We call this problem the *influence limitation problem.*

In this context, we consider the social network as an edge-weighted graph. The nodes represent users of the network and edges represent a relationship between these users. Edges can be undirected, indicating that the relationship is symmetric (users are friends with each other), or they can be directed, indicating one-way communication such as a publisher/subscriber relationship (follower relationship in Twitter). We use edge weights to quantify the "influence" that one node has upon another node; the weight of edge $e_{i,j}$ is an indicator of the likelihood that node j will agree with and pass on information it receives from node i. Nodes that have adopted the idea or information are called *active* and those that have not are *inactive.*

We consider two different information diffusion models, both similar to the Independent Cascade Model (ICM). In ICM, the information diffusion process takes place in discrete rounds. When node i is first activated in round r, it tries to activate each inactive neighbor j; it succeeds with probability $p_(i, j)$. Whether or not it succeeds, node i does not try to activate anyone in subsequent rounds. In order to capture simultaneous spread of two cascades (the initial misinformation campaign and the limiting campaign), we introduced two extensions to ICM called the Multi-Campaign Independent Cascade Model (MCICM) and Campaign-Oblivious Independent Cascade Model (COICM) [BAEA11a]. We omit the details of these models due to space limitations but note that they are similar to a number of other models in literature [BKS07, DGM06, CNWvZ07, KOW08].

Our objective is to minimize the number of people affected by the misinformation campaign by "vaccinating" users through a *limiting campaign.* Let the campaign that is to be minimized be campaign C and the initial set of nodes activated in campaign C be A_C. The limiting campaign is called campaign L and the initial activation set is A_L. For simplicity, we assume that a user can be active in only one campaign, *i.e.*, once the user has decided on a side, he will not change his mind. Given a network and a diffusion model, suppose that a campaign that is spreading bad information is detected r rounds after it began. Given a budget l, select l individuals for initial activation with the competing information such that the expected number of nodes eventually activated in campaign C is minimized. Let $IF(A_C)$ denote the influence set of campaign C without the presence of campaign L, *i.e* the set of nodes that would accept campaign C if there were no limiting campaign. We define the function $\pi(A_L)$ to be the size of the subset of $IF(A_C)$ that campaign L prevents from adopting campaign C. Then, the influence limitation problem is equivalent to selecting A_L such that the expectation of $\pi(A_L)$ is maximized.

```
Initialize $A_L$ to $\emptyset$
for $i = 1$ to $l$ do
    Choose node $i$ that maximizes $\pi(A_L \cup \{i\}) - \pi(A_L)$;
    Set $A_L \leftarrow A_L \cup \{i\}$;
```

Fig. 1. Greedy algorithm to select the set for initial activation in the limiting campaign

We now outline a potential solution to a simplified version of this problem. We assume that there is only a single source of information for campaign C, meaning $|A_C| = 1$ and that information diffusion follows the Multi-Campaign Independent Cascade Model. Finally, we assume that L is accepted by users with probability 1 (it may be much easier to convince a user of the truth than a falsehood). We refer to this notion as *high-effectiveness property.* In a recent study [BAEA11a], we have shown that $\pi(A_L)$ is a monotone, sub-modular function for this setting. We have also proved the same result in the context of Campaign-Oblivious Independent Cascade Model, even without the *high-effectiveness property.* Therefore, for both diffusion models, the greedy algorithm given in Figure 1 yields an A_L for which $\pi(A_L)$ is within $1 - 1/e$ of optimal [BAEA11a].

3 Opinion Dynamics

Another process that is of interest in social networks research is the process by which a group of individuals in a social network form an opinion about a piece of information or an idea and how interpersonal influence affects the opinion formation process. This process is known as *opinion dynamics.* While this process shares some similarities with the diffusion of an information campaign, it differs in the main respect that the individual opinions evolve over time, continuously changing in response to interactions with other individuals and possibly external inputs.

Research on opinion formation in social groups predates the advent of online social networks by decades, and several formal mathematical models of the opinion formation process have been proposed [Fre56, DeG74, Leh75, FJ99, HK02]. These early works are based on the assumption that all individuals interact with all friends simultaneously in a synchronized fashion. In online social networks, however, individuals are spread out across space and time and they interact with different communities and friends at different times and with different frequencies. We investigate how these previously proposed models of opinion dynamics can be augmented to incorporate the asynchronous nature of social interactions that arise in online social networks.

We first briefly review the general opinion dynamics model. The social network is modeled as a graph $G = (V, E)$ where the vertices represent users or *agents*, with $|V| = n$, and the edges represent relationships between users, with

$|E| = m$. The graph may be directed or undirected. A directed graph is used to model networks where relationships are not symmetric, for example, the follower relationship in Twitter. An undirected graph models a network with symmetric relationships such as the friend relationship in Facebook. We say that agent i is a *neighbor* of agent j if $(i, j) \in E$. Each individual i has an initial opinion $x_i(0)$. The opinion is assumed to be a real number, for example a numerical representation of the individual's support for an issue. The agreement process takes place in discrete rounds. In each round, each individual updates his opinion based on information exchanged along edges in the network graph, taking a weighted average of his own opinion and the opinions of his neighbors. Let w_{ij} be the weight that agent i places on the opinion of agent j with the normalization requirement that $\sum_{j=1}^{n} w_{ij} = 1$. In each round, individual i updates his opinion as follows:

$$x_i(t + 1) = w_{i1}x_1(t) + w_{i2}x_2(t) + \ldots + w_{in}x_n(t),$$

where $w_{ij} > 0$ only if $(i, j) \in E$. We note that this formulation admits the possibility that $w_{ij} = 0$ even if $(i, j) \in E$, meaning i does not place any weight on the opinion of j even though they are neighbors in the network. The evolution of all of the opinions, called the *opinion profile*, is captured by the following recursion:

$$x(t + 1) = W(t, x(t))x(t).$$

$x(t)$ is the n-vector of opinions at round t, and $W(t, x(t))$ is an $n \times n$ matrix that gives the edge weights, the interpersonal influence, for round t. This general model allows for the possibility that these edge weights may change over time and may depend on the current opinion profile.

In the following sections, we restrict our discussion to the *classical model* of opinion dynamics that was proposed by De Groot [DeG74] and Lehrer [Leh75], where the edge weights are assumed to remain constant throughout the opinion formation process. The process is given by the recursion

$$x(t + 1) = Wx(t) \tag{1}$$

where W is an $n \times n$ matrix.

Due to the simple form of this model, it is possible to analyze properties of the opinion formation process and predict the outcome by examining the W matrix. In particular, it can be shown that (1) Individuals reach agreement if and only if $|\lambda_2(W)| < 1$, where $\lambda_2(W)$ is the second largest eigenvalue of W by magnitude; (2) If $|\lambda_2(W)| < 1$ and W is symmetric, the consensus value is the average of the initial opinions. If W is not symmetric, the consensus value is a weighted average of the initial opinions; and (3) If agreement will occur, the number of rounds required for individuals to be within ϵ of the consensus value is $\log \epsilon / \log(\lambda_2(W))$. We next describe several extensions to the classical opinion dynamics that capture the types of social interactions exhibited in online networks.

It has been observed that users interact frequently with only a small subset of their neighbors and that communication is infrequent along many edges

[WBS$^+$09]. It is reasonable to expect that the frequency of interaction will have a large impact on the evolution of opinions in online social networks. To capture the notion of interaction frequency, we associate a (unique) probability of communication p_{ij} with each edge $(i, j) \in E$. The value p_{ij} is the probability that agents' i and j will communicate in each round. The evolution of the opinion profile with probabilistic interactions is given by the following recursion,

$$x(t+1) = \left(I - \sum_{(i,j)\in E} \delta_{ij}(t) w_{ij} L_{ij} \right) x(t) \tag{2}$$

where $\delta_{ij}(t)$ are independent Bernoulli random variables with

$$\delta_{ij}(t) \ := \ \begin{cases} 1 & \text{with probability } p_{ij} \\ 0 & \text{with probability } 1 - p_{ij}. \end{cases}$$

This model has been adapted from the model for multi-agent consensus in stochastic networks [PBE10]. Each L_{ij} is the weighted Laplacian matrix of the graph $G_{ij} = (V, \{(i, j)\})$, the graph that contains the same n vertices as the original graph G and the single edge (i, j). When $\delta_{ij} = 1$, agents i and j exchange information just as they did in the classical model. When $\delta_{ij} = 0$, agents i and j do not communicate.

In our recent work [PB10, PBE10], we have derived a matrix-valued, Lyapunov-like operator $\mathcal{W}(\cdot)$ that describes the evolution of the covariance of the opinion profile in this stochastic model of opinion dynamics and we have provided analysis similar to the well-known results for classical opinion dynamics.

4 Information Trend Analysis

Social networks provide a large-scale information infrastructure for people to discuss and exchange ideas about variety of topics. Detecting trends of such topics is of significant interest for many reasons. For one, it can be used to detect emergent behavior in the network, for instance a sudden increase in the number of people talking about explosives or biological warfare. Information trends can also be viewed as a reflection of societal concerns or even as a consensus of collective decision making. Understanding how a community *decides* that a topic is trendy can help us better understand how ad-hoc communities are formed and how decisions are made in such communities. In general, constructing "useful" trend definitions and providing scalable solutions that detect such trends will contribute towards a better understanding of human interactions in the context of social media.

Before we study the problem of finding trendy topics in a social network, we first need to develop a clear definition of "trendiness". Assume users of a network can choose to (or not to) broadcast their opinions about various topics at any point in time. Assume further that we can abstract away what the topic is from what a user broadcasts. A simple definition of trendy topics can be the frequent items throughout the entire history of user broadcasts. The problem, defined this way, is simply to find the frequent items in a stream of data, also referred to as *heavy hitters*. The *frequent elements problem* is well studied and several scalable, online solutions have been proposed [CCFC02, CM05, MAE06, MM02, DLOM02]. While the heavy hitters view trend definition is compelling because of the existence of scalable algorithms, this simple definition overlooks various important aspects such as the spatio-temporal dimensions of human interaction and the network structure and its effect on the emergence of trends. In the following, we explore structural trend analysis as an example that depend on structural connections between the users who are broadcasting. Information trends at the granularity of spatio-temporal dimensions remains future work.

Assuming that information diffusion on a social network is a substantial part of the process that creates the information trends, properties that are defined in the context of the network structure are of significant interest. For example, consider a group of friends in a large social network like Facebook discussing an attack. Detecting this new interest of this specific group on "attacks" can be of great importance. Note that especially for cyber attacks, those people do not necessarily need to be in the same geographical region, and in some instances, this geographic information is not even available. In essence, a structural trend is a topic that is identified as "hot" within structural subgroups of the network. The challenges are to formally define the notions of a structural subgroup and to develop techniques to detect *structural trends.*

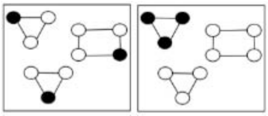

Fig. 2. Black nodes represent nodes talking about topic x, whereas white nodes represent the nodes that are not

As a starting point, we consider the problem of identifying the number of connected pairs of users in a social network that are discussing a specific topic. We refer to this as detecting *correlated trends.* Consider the two graphs in Figure 2. The black nodes correspond to people that are talking about a specific topic and white nodes are people who are not talking about it. Even though both graphs have the same number of people talking about this topic, in the graph on the right, the people talking about the issue are a part of a more clustered subgraph, giving the topic a higher structural significance. The number of pairs of users talking about the topic in the graph on the left is 0 whereas the pair count is 3 for the graph on the right. The detection of structural trends is harder to solve than the traditional definition of trends since in the traditional setting a counter can be updated on each new data item without dependence on any future or past items. This is not the case for the correlated trends. Value addition of a tuple $\langle n_i, T_j \rangle$ to topic T_j, where n_i represents a node in the graph and T_j is an arbitrary topic depends on other tuples involving T_j and neighbors of n_i. The detection of structural trends calls for new, graph-oriented solutions.

The exact solution for *correlated trends* that requires keeping track of all topics for all nodes is not scalable for large networks so we need to explore approximation algorithms. Here we describe possible solutions: (i) Use the activity level and degree of a node as a heuristic to limit the number of nodes monitored per topic; (ii) Only monitor those topics that are frequent in the traditional sense and for such topics find the order of their importance w.r.t. correlated trendiness; (iii) As demonstrated by Mislove et al. [MMG$^+$07], there is a small subset of nodes in social networks without which the network would be a collection of many small disconnected subgraphs. This property can be exploited to partition the graph into smaller sub-graphs and apply the counting algorithms in parallel. Query processing requires periodically merging the counts from multiple sub-partitions with the counts from highly connected nodes. This approach is highly scalable in the MapReduce paradigm [DG08].

Another solution we propose to evaluate involves a semi-streaming approximation algorithm. We essentially reduce the problem of evaluating the importance of each topic with respect to the correlated trendiness notion to a problem of counting local triangles in a graph, *i.e.* counting the number of triangles incident to every node $n \in N$ in the graph. Consider a social network graph $G = (N, E)$, a set of all topics T and stream of node-topic tuples S, where each tuple is in the form: $\langle n_i, T_j \rangle$ s.t. $n_i \in N$ and $T_j \in T$. Let us create a graph $G' = (N', E')$ s.t. $N' = N \cup T$ and $E' = \{(u, v)|(u, v) \in E \wedge (u, v) \in S\}$. The number of connected pairs of users in the social network G that are discussing a specific topic T_j is simply the number of triangles incident to node T_j in G'. Using approximation algorithms based on the idea of min-wise independent permutations similar to [BBCG08], we are able to provide a solution using O($|N'|$) space in main memory and performing O(log$|N'|$) sequential passes over E'. Note that the naive solution would need O($|N| \cdot |T|$) space in main memory.

Alternatively, we can define a structural trend be the other extreme, where we are interested in the number of *unrelated* people interested in a specific topic and in trends that results from these unrelated people. We call these *uncorrelated trends*. Going back to our example of two graphs in Figure 2, for the graph on the left this count will be 3, whereas it will be only 1 for the graph on the right. This definition of trendiness can be used capture the notion of the *trustworthiness* of a trend. In this case trendiness of a topic will not be biased by a discussion amongst a small clustered group. We note that we have only considered two extremes of structural trends. Identifying alternative definitions of structural trends that will span the entire spectrum remains future work [BAEA11b].

5 Concluding Remarks

Internet technologies and online social networks are changing the nature of social interactions and user behaviors. Recent evidence indicates that 45% of users in the U.S. state that the Internet played a crucial or important role in at least one major decision in their lives in the last two years, such as attaining additional career training, helping themselves or someone else with a major illness or medical condition, or making a major investment or financial decision [HR06]. The significant role of online social networks in human interactions motivates our goals of generating a greater understanding of social interactions in online networks through data analysis, the development of reliable models that can predict outcomes of social processes, and ultimately the creation of applications that can shape the outcome of these processes. In this paper, we have presented a preliminary formulation of modeling and analyzing *information diffusion*, *opinion dynamics*, and *information trends* in online social networks.

Acknowledgments. This work was partially supported by the NSF under grant IIS-1135389.

References

[AG05] Adamic, L.A., Glance, N.: The political blogosphere and the 2004 u.s. election: divided they blog. In: LinkKDD 2005: Proceedings of the 3rd International Workshop on Link Discovery, pp. 36–43 (2005)

[BAEA10] Budak, C., Agrawal, D., El Abbadi, A.: Where the blogs tip: connectors, mavens, salesmen and translators of the blogosphere. In: SIGKDD Workshop on Social Media Analytics (2010)

[BAEA11a] Budak, C., Agrawal, D., El Abbadi, A.: Limiting the spread of misinformation in social networks. In: Proceedings of the 20th International Conference on World Wide Web, WWW 2011, pp. 665–674 (2011)

[BAEA11b] Budak, C., Agrawal, D., El Abbadi, A.: Structural trend analysis for online social networks. In: VLDB 2011 (2011)

[BBCG08] Becchetti, L., Boldi, P., Castillo, C., Gionis, A.: Efficient semi-streaming algorithms for local triangle counting in massive graphs. In: KDD 2008: Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 16–24. ACM, New York (2008)

[BGMI05] Baumes, J., Goldberg, M., Magdon-Ismail, M.: Efficient identification of overlapping communities. In: IEEE International Conference on Intelligence and Security Informatics (ISI), pp. 27–36 (2005)

[BKS07] Bharathi, S., Kempe, D., Salek, M.: Competitive influence maximization in social networks. In: Deng, X., Graham, F.C. (eds.) WINE 2007. LNCS, vol. 4858, pp. 306–311. Springer, Heidelberg (2007)

[CCFC02] Charikar, M., Chen, K., Farach-Colton, M.: Finding frequent items in data streams. In: Widmayer, P., Triguero, F., Morales, R., Hennessy, M., Eidenbenz, S., Conejo, R. (eds.) ICALP 2002. LNCS, vol. 2380, pp. 693–703. Springer, Heidelberg (2002)

[Cla05] Clauset, A.: Finding local community structure in networks. Physical Review E (Statistical, Nonlinear, and Soft Matter Physics) 72(2), 026132 (2005)

[CM05] Cormode, G., Muthukrishnan, S.: What's Hot and What's Not: Tracking Most Frequent Items Dynamically. ACM Trans. Database Syst. 30(1), 249–278 (2005)

[CNWvZ07] Carnes, T., Nagarajan, C., Wild, S.M., van Zuylen, A.: Maximizing influence in a competitive social network: a follower's perspective. In: ICEC 2007: Proceedings of the Ninth International Conference on Electronic Commerce, pp. 351–360 (2007)

[CYY09] Chen, W., Wang, Y., Yang, S.: Efficient influence maximization in social networks. In: Proceedings of the 15th ACM International Conference on Knowledge Discovery and Data Mining, pp. 199–208 (2009)

[DeG74] DeGroot, M.H.: Reaching a consensus. Journal of the American Statistical Association 69, 118–121 (1974)

[DG08] Dean, J., Ghemawat, S.: Mapreduce: simplified data processing on large clusters. ACM Commun. 51(1), 107–113 (2008)

[DGM06] Dubey, P., Garg, R., De Meyer, B.: Competing for customers in a social network. Cowles Foundation Discussion Papers 1591, Cowles Foundation, Yale University (November 2006)

[DLOM02] Demaine, E.D., López-Ortiz, A., Munro, J.I.: Frequency estimation of internet packet streams with limited space. In: Möhring, R.H., Raman, R. (eds.) ESA 2002. LNCS, vol. 2461, pp. 348–360. Springer, Heidelberg (2002)

[DR01] Domingos, P., Richardson, M.: Mining the network value of customers. In: Proceedings of the 7th ACM International Conference on Knowledge Discovery and Data Mining, pp. 57–66 (2001)

[Far09] Farrell, M.B.: Schwarzenegger tweets about swine flu. so does everyone else. Christian Science Monitor (April 2009)

[FJ99] Friedkin, N.E., Johnsen, E.C.: Social influence networks and opinion change. Advances in Group Processes 16, 1–29 (1999)

[Fre56] French, J.R.P.: A formal theory of social power. Psychological Review 63, 181–194 (1956)

[GL08] Ghosh, R., Lerman, K.: Community Detection Using a Measure of Global Influence. In: Giles, L., Smith, M., Yen, J., Zhang, H. (eds.) SNAKDD 2008. LNCS, vol. 5498, pp. 20–35. Springer, Heidelberg (2010)

[Gla02] Gladwell, M.: The Tipping Point: How Little Things Can Make a Big Difference. Back Bay Books (January 2002)

[GLM01] Goldenberg, J., Libai, B., Muller, E.: Talk of the network: A complex systems look at the underlying process of word-of-mouth. Marketing Letters 12(3), 209–221 (2001)

[Gra78] Granovetter, M.: Threshold models of collective behavior. American Journal of Sociology 83(6), 1420–1443 (1978)

[Gre07] Gregory, S.: An algorithm to find overlapping community structure in networks. In: Kok, J.N., Koronacki, J., Lopez de Mantaras, R., Matwin, S., Mladenič, D., Skowron, A. (eds.) PKDD 2007. LNCS (LNAI), vol. 4702, pp. 91–102. Springer, Heidelberg (2007)

[Gro09] Grossman, L.: Iran protests: Twitter, the medium of the movement. Time (online) (June 2009)

[HK02] Hegselmann, R., Krause, U.: Opinion dynamics and bounded confidence models, analysis, and simulation. Journal of Artificial Societies and Social Simulation 5(3) (2002)

[HP09] Hughes, A.L., Palen, L.: Twitter adoption and use in mass convergence and emergency events. In: Proceedings of the 6th International Information Systems for Crisis Response and Management Conference (2009)

[HR06] Horrigan, J., Rainie, L.: When facing a tough decision, 60 million americans now seek the internet's help: The internet's growing role in life's major moments (2006), http://pewresearch.org/obdeck/?ObDeckID=19 (retrieved October 13, 2006)

[KKT03] Kempe, D., Kleinberg, J.M., Tardos, É.: Maximizing the spread of influence through a social network. In: Proceedings of the Ninth ACM International Conference on Knowledge Discovery and Data Mining, pp. 137–146 (2003)

[KOW08] Kostka, J., Oswald, Y.A., Wattenhofer, R.: Word of Mouth: Rumor Dissemination in Social Networks. In: 15th International Colloquium on Structural Information and Communication Complexity (SIROCCO) (June 2008)

[KSNM09] Kimura, M., Saito, K., Nakano, R., Motoda, H.: Finding Influential Nodes in a Social Network from Information Diffusion Data. In: Social Computing and Behavioral Modeling. Springer, US (2009)

[Leh75] Lehrer, K.: Social consensus and rational agnoiology. Synthese 31(1), 141–160 (1975)

[LKG$^+$07] Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., Glance, N.: Cost-effective outbreak detection in networks. In: Proceedings of the 13th ACM International Conference on Knowledge Discovery and Data Mining, pp. 420–429 (2007)

[MAE06] Metwally, A., Agrawal, D., El Abbadi, A.: An integrated efficient solution for computing frequent and top-k elements in data streams. ACM Trans. Database Syst. 31(3), 1095–1133 (2006)

[MM02] Manku, G.S., Motwani, R.: Approximate frequency counts over data streams. In: Proc. 28th Int. Conf. on Very Large Data Bases, pp. 346–357 (2002)

[MMG$^+$07] Mislove, A., Marcon, M., Gummadi, K.P., Druschel, P., Bhattacharjee, B.: Measurement and analysis of online social networks. In: Proc. 7th ACM SIGCOMM Conf. on Internet Measurement, pp. 29–42 (2007)

[New03] Newman, M.E.J.: Fast algorithm for detecting community structure in networks. Physical Review E 69 (September 2003)

[New06] Newman, M.E.J.: Modularity and community structure in networks. Proceedings of the National Academy of Sciences 103(23), 8577–8582 (2006)

[PB10] Patterson, S., Bamieh, B.: Interaction-driven opinion dynamics in online social networks. In: SIGKDD Workshop on Social Media Analytics (2010)

[PBE10] Patterson, S., Bamieh, B., El Abbadi, A.: Convergence rates of distributed average consensus with stochastic link failures. IEEE Transactions on Automatic Control 55(4), 880–892 (2010)

[PSL90] Pothen, A., Simon, H.D., Liou, K.-P.: Partitioning sparse matrices with eigenvectors of graphs. SIAM J. Matrix Anal. Appl. 11(3), 430–452 (1990)

[RD02] Richardson, M., Domingos, P.: Mining knowledge-sharing sites for viral marketing. In: Proceedings of the 8th ACM International Conference on Knowledge Discovery and Data Mining, pp. 61–70 (2002)

[WBS$^+$09] Wilson, C., Boe, B., Sala, A., Puttaswamy, K.P.N., Zhao, B.Y.: User interactions in social networks and their implications. In: Proc. 4th ACM European Conference on Computer Systems, pp. 205–218 (2009)

[WCSX10] Wang, Y., Cong, G., Song, G., Xie, K.: Community-based greedy algorithm for mining top-k influential nodes in mobile social networks. In: Proceedings of the 16th ACM International Conference on Knowledge Discovery and Data Mining (2010)

[WF94] Wasserman, S., Faust, K.: Social Network Analysis. Cambridge University Press, Cambridge (1994)

[ZWZ07] Zhang, S., Wang, R.S., Zhang, X.S.: Identification of overlapping community structure in complex networks using fuzzy c-means clustering. Physica A: Statistical Mechanics and its Applications 374(1), 483–490 (2007)