Structural Trend Analysis for Online Social Networks*

Ceren Budak
Department of Computer Science, UCSB
Santa Barbara, USA
cbudak@cs.ucsb.edu

Divyakant Agrawal
Department of Computer Science, UCSB
Santa Barbara, USA
agrawal@cs.ucsb.edu

Amr El Abbadi
Department of Computer Science, UCSB
Santa Barbara, USA
amr@cs.ucsb.edu

ABSTRACT

The identification of popular and important topics discussed in social networks is crucial for a better understanding of societal concerns. It is also useful for users to stay on top of trends without having to sift through vast amounts of shared information. Trend detection methods introduced so far have not used the network topology and has thus not been able to distinguish viral topics from topics that are diffused mostly through the news media. To address this gap, we propose two novel structural trend definitions we call *coordinated* and *uncoordinated* trends that use friendship information to identify topics that are discussed among clustered and distributed users respectively. Our analyses and experiments show that structural trends are significantly different from traditional trends and provide new insights into the way people share information online. We also propose a sampling technique for structural trend detection and prove that the solution yields in a gain in efficiency and is within an acceptable error bound. Experiments performed on a Twitter data set of 41.7 million nodes and 417 million posts show that even with a sampling rate of 0.005, the *average precision* is 0.93 for *coordinated* trends and 1 for *uncoordinated* trends.

1. INTRODUCTION

Social networks provide large-scale information infrastructures for people to discuss and exchange ideas about different topics. Detecting trends of such topics is of significant interest for many reasons. For one, trends can be used to detect emergent or suspicious behavior in the network. They 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 detection methods for such definitions will contribute towards a better understanding of interactions in the context of social media.

Trends in social networks have recently been a major focus of interest among researchers studying them from perspectives such as temporal [24] and geographical [32, 30] dimensions. A similar interest can be observed in industry [36, 19]. Furthermore, the importance society places on trends in social networks is increasing. For instance, Twitter trends [36] have been a testament to societal concerns to such an extend that a possible exclusion of the hashtag #wikileaks from the trends list in Twitter created a large dispute which the company had to officially address [40].

Although trends in social networks have been extensively studied, to our knowledge all the published work in this area ignores the structural properties of the social network that created these trends. In today's social networks where users are highly influenced by their friends, trend definitions that reach beyond simple heavy-hitters approaches to integrate the importance of such flow of influence can be of great benefit. The main purpose of this paper is to introduce social network structure into trend analysis, propose two novel structure-based trend definitions, emphasize their significance and provide efficient online solutions for them. Since information diffusion is a substantial part of the process that creates the information trends, properties that are defined in this context are of significant interest. Although current trend definitions used by the industry [36] are good at detecting trends at global scale, their shortcomings such as their vulnerability to spammers or inability to detect interesting activity in different communities make them less valuable from an analytical perspective [37]. The new trend definitions introduced in this paper provide methods for a deeper analysis of activity in social networks.

A structural trend is a topic that is popular 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*. As a starting point, we focus on identifying trends where the trendiness of a topic can be characterized by the number of connected pairs of users discussing it. We refer to this as detecting *coordinated trends*. This notion favors topics that are discussed among clustered nodes in the network. Secondly, we study another notion of trendiness that we call *uncoordinated trends* where the score of a topic is based on the number of *unrelated* people interested in it. This definition of trendiness, not being biased by a discussion amongst a small clustered group, can be used to capture the notion of the *trustworthiness* of a trend. We establish the value of these two novel trend definitions by identifying the types of topics they detect that would be undetected using traditional trend detection methods. We also provide graph-oriented solutions for detection of structural trends. Considering both the large scale of social networks and the sheer volume of information shared, we introduce a sampling based technique that provides efficiency while still remaining within an acceptable error bound.

To the best of our knowledge, this is the first work that incorporates the structure of a graph to the definition of trends. We introduce two novel trend definitions based on the structure of the net-

*This work is partially supported by NSF Grant IIS-1135389.

Proceedings of the VLDB Endowment, Vol. 4, No. 10
Copyright 2011 VLDB Endowment 2150-8097/11/07... $ 10.00.

work. We experimentally and analytically show that *coordinated* trends are significantly different from *traditional* trends whereas the difference for *uncoordinated* trends is less pronounced. We show that *structural* trends identify interesting activities in social networks. In Section 2 we list related work. In Section 3 we formally define *structural* trends and demonstrate their significance in Section 4. Section 5 introduces and experimentally studies the accuracy and efficiency of sampling-based solutions showing that a high average precision of 0.93 can be achieved even with a sampling probability of 0.005. Finally Section 6 concludes the paper.

2. RELATED WORK

Trends in social networks have recently been a focus of interest for many researchers. A bulk of research concentrated on trends from a temporal point of view [2, 17, 22, 24]. Kwak et al. [22] study and compare trending topics in Twitter reported by Twitter [36] with those in other media. The results show that the majority of topics are headline news or persistent news in nature. Leskovec et al. [24] also study temporal properties of information shared in social networks by tracking "memes" across the blogosphere. Another important characteristic of news or discussions in social networks is the spatial properties of the agents that are involved in the discussion or the source of the news. A recent work by Teitler et al. [32] collects, analyzes, and displays news stories on a map interface. A follow-up study performs similar techniques to identify geographical information in news in Twitter [30]. Although trend analysis based on temporal and spatial characteristics is important for a better understanding of trends, they are orthogonal to the approaches introduced in this study. Unlike earlier studies, we focus on structural properties of the network that create trends.

Trends, in the traditional sense, can simply be defined as the frequently mentioned topics throughout the stream of user activities. This problem is simply to find the frequent items in a stream, also referred to as *heavy hitters*. The *frequent elements problem* is well studied and several scalable, online solutions have been proposed. A survey of such methods can be found in [14]. Unlike these techniques, our solution is not oblivious to the graph structure. Several works have studied structural properties of graphs in a streaming or semi-streaming fashion. One problem that is relevant to our work is counting triangles in a graph stream. There are three types of solutions for this problem: exact counting [4], streaming [5, 12, 20] and semi-streaming algorithms [6, 34]. Although streaming algorithms [5, 12, 20] provide efficient solutions, they solve the global triangle counting problem, which counts *all* the triangles in a graph whereas structural trendiness requires solutions closer to local triangle counting. In that sense, problems studied in [6, 34] are closer to the problem studied in this paper. However, these works provide a semi-streaming solution and therefore are not applicable as an online solution. Structural properties of graphs also have significance for research in the area of influence spread [21, 24, 13, 8, 11, 9, 10]. Different from these works, we focus on large-scale data analysis.

In a recent work that was executed concurrently with our work, Agrawal et al. [1] posit that the nature of information items plays a vital role in information spread and propose a model that assigns to every information item two parameters: endogeneity and exogeneity. Our model is similar to this model and is more general since different endogeneity and exogeneity values can be assigned to nodes as well as topics. Their findings support the claims put forward in our work by showing that there are topics that are inherently of different nature. The approaches presented in [1] and our work are complimentary since [1] develops a maximum-likelihood framework for estimating parameters of the model while we develop online solutions for detecting endogenous and exogenous trends.

3. PROBLEM DEFINITION

Consider a directed graph $G = (N, E)$ representing a social network consisting of nodes N and edges E where $e_{j,i} \in E$ means node n_i is a neighbor of n_j. At any point nodes of the network can share information on any topic with their *neighbors*. We model each such *mention* by node n_i on a specific topic T_x as a tuple $\langle n_i, T_x \rangle$. We refer to the history of such tuples as *stream* and denote it using S. Note that, topic extraction is a hard problem in its own right and we will not focus on this problem in this paper. Under this model, *traditional* trendiness of T_x can be defined as the total number of times T_x is mentioned as:

$$f(T_x) = \sum_{n_i \in N} C_{i,x} \tag{1}$$

where $C_{i,x}$ represents the number of mentions of the form $\langle n_i, T_x \rangle$ in S. This trend definition is oblivious to the structure of the network. We propose two new alternative trend definitions, namely *coordinated* and *uncoordinated* trends, to capture this characteristic. Our main goal for the first trend definition is to give a high score to topics that are discussed heavily in a cluster of tightly connected nodes. Therefore, we introduce *coordinated* trendiness score as follows:

$$g(T_x) = \sum_{n_i \in N} (C_{i,x} \sum_{n_k \in N_i} C_{k,x}) \tag{2}$$

where $N_i = \{n_k | e_{i,k} \in E\}$. Informally, function $g(T_x)$ counts the number of pairs of mentions of neighboring nodes. This is achieved by weighing the traditional count for each node by the sum of the counts for all its neighbors. There are four characteristics of Equation 2 that are of interest from an analytical point of view:

(i) It assigns a high score to those topics that are discussed by a large number of pairs of connected nodes. Consider the two graphs in Figure 1 where the black nodes correspond to nodes that mention topic T_x and white nodes do not. Even though $f(T_x)$ has the same value for both graphs, in the graph on the right, the nodes mentioning T_x are a part of a clustered subgraph, giving T_x a higher structural significance. Capturing this notion, $g(T_x) = 6$ for the graph on the right and $g(T_x) = 0$ for the graph on the left.

(ii) It assigns a high score to topics with a large number of *mentions* by using a count of *mentions* per node rather than simply counting number of pairs of nodes.

(iii) By multiplying the counts of *mentions* of neighbors, g favors a uniform distribution of mentions per node in the case of a complete graph where each node, having the same degree centrality, is of same significance. Consider two topics T_x and T_y, each having $f = 2N$ where N is the number of nodes. For T_x assume each node has 2 mentions while for T_y , the first node has $N + 1$ mentions and the other nodes have 1 mention each. In this case since $g(T_y) = 3N(N-1)$ while $g(T_x) = 4N(N-1)$, the score of T_x with a uniform distribution has a higher score.

(iv) In a power law graph with a small number of highly connected and influential nodes, it is desirable that a score function is biased towards mentions from *influential* nodes. As the counting scheme for Equation 2 is a multiplication of mentions over edges, for nodes that have a large number of edges, this results in a large number. Consider a graph with a one-level tree structure where each edge is bidirectional, a topic T_x with K mentions from the root of the tree and 1 mention from the rest $N-1$ has score $2K(N-1)$, while a topic T_y with K mentions from one of the leaves and 1 mention from each other node only has a score of $2K + 2N - 4$. For any $N > 2$, T_x has a higher score than T_y. Note that, we use degree centrality as a notion of *influence*, an idea used in the literature [9].

This new paradigm, in addition to addressing these four points, captures *all possible forms of influence propagation between any two neighbors*. Consider a stream: $...m_1{:}\langle n_1, T_x \rangle, ..., m_2{:}\langle n_1, T_x \rangle,$

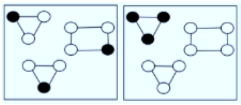


Figure 1: Black nodes represent nodes that mention topic T_x, whereas white nodes represent the nodes that do not.

m_3:$\langle n_2, T_x\rangle$,...,m_4:$\langle n_2, T_x\rangle$,...,m_5:$\langle n_1, T_x\rangle$... where n_1 and n_2 are neighbors. In this setting n_1 sharing m_1 and m_2 *might have influenced* n_2 to share m_3 and m_4. Similarly m_3 and m_4 might have influenced n_1 to share m_5. Equation 2 captures all the pairs of *possible flow of influence* for this example and in general for undirected graphs. For directed graphs, although the *coordinated trend score* does not correspond directly to this notion, it captures a similar behavior. Our main goal for the second trend definition is to give high score to topics that are discussed heavily by unconnected nodes, giving the topic a general credibility. Therefore, we introduce *uncoordinated* trendiness score as follows:

$$h(T_x) = \sum_{n_i \in N} (C_{i,x} \sum_{n_k \in N \setminus (n_i \cup N_i)} C_{k,x}) \tag{3}$$

Informally, function $h(T_x)$ counts the number of pairs of mentions by unconnected nodes. Going back to our example in Figure 1, for the graph on the left $h(T_x) = 6$, whereas $h(T_x) = 0$ for the graph on the right. Due to space limitations, we omit detailed analysis of characteristics of the $h(T_x)$. However we note that it favors topics with a large number of total *mentions*, while downgrading the importance of topics that are mentioned by a small set of connected nodes. Therefore *uncoordinated* trends can be used to capture the notion of dispersed and widespread interest and could be used to reflect the *trustworthiness* of a topic.

We denote *top-k* topics w.r.t. f, g, h scores as *traditional, coordinated* and *uncoordinated* trends respectively. The combined class of *coordinated* and *uncoordinated* trends is referred to as *structural trends*. In the following sections, we will demonstrate the usefulness of *structural* trends and provide solutions for detecting them.

4. STRUCTURAL TRENDS SIGNIFICANCE

In this section, we will demonstrate the value of structural trends by identifying the interesting activity detected using such new trend definitions. We will demonstrate the significance of Equations 2 and 3 by addressing the following questions: 1) Are the *structural* trends different from *traditional* trends? 2) What is the nature of topics detected using structural trends? We make use of two different methods to answer these questions. First, we develop a model of diffusion of an arbitrary number of information campaigns in a social network. The importance of structural trends is then identified with respect to the parameters of this model. Second, we analyze data from Twitter, a large-scale online social network and identify the types of topics detected using structural trends and focus on their significance.

4.1 Model-Based Validation

In order to systematically evaluate the significance of *structural trends*, we need to model the process that creates trends in a social network and identify characteristics of social networks or topics that validate the significance of structural trends. Although there are a number of models of diffusion of a single information campaign [21], there is little research on modeling of concurrent information campaigns with the exception of [11, 8, 13] which study the diffusion of *two* concurrent campaigns. Here we introduce a natural extension of the widely used Independent Cascade model [21] that models the diffusion of an arbitrary number of campaigns. We call this model the *Independent Trend Formation Model (ITFM)* as the diffusion of topics are assumed to be independent of each other.

ITFM captures nodes as entities that are influenced by their neighbors as well as external entities such as news media. We model a social network as a directed graph. There are a set of m topics $T = \{T_1, ..., T_m\}$. Information diffusion proceeds in discrete time steps. At each step, nodes *mention* zero or more topics. As we would like to model the different types of influence, we assign two types of probabilities to each node n_i: $p_{i,x}$ and $q_{i,j,x}$ that denote the probability that n_i will *mention* T_x independently from any of its neighbors (external influence such as news media) and the probability that n_i will *mention* T_x that its neighbor n_j *mentioned* in the earlier discrete time step (peer influence). If for a topic T_x the p probabilities are high, T_x spreads mostly through the news media channels, whereas if the q probabilities are high, this means T_x is *viral*, spreading through peer influence.

Using *ITFM* as the information diffusion model, we performed experiments on synthetic power-law graphs since social networks have power-law degree distribution [29]. In order to produce the synthetic graphs, we used the *Nearest Neighbor* model as it is shown to accurately capture various statistical metrics of real social networks [29]. There are two important parameters for the Nearest Neighbor Model; u, the probability two nodes with a distance of two are connected at a time step and k, the number of pairs of existing nodes connected at a time step. We set $u = 0.8$ and $k = 1$ since these settings fit a real social network, namely the Facebook Monterey Bay Network [29]. The experiments in this section were performed on a 500 node power-law graph with a set of 50 topics.

Question 1: Are the *structural* trends different from *traditional* trends? Do structural trends provide extra information that could not be obtained otherwise? Or in other words, *how similar are structural and traditional trends?* To measure similarity between *traditional* and *coordinated* trends, we used *Spearman rank correlation coefficient (SRCC)* [26]:

$$\rho_{trad-coor} = 1 - \frac{6\sum {d_x}^2}{n(n^2-1)} \tag{4}$$

where d_x is the difference between the ranks of topic T_x under *coordinated* and *traditional* trends. Similarly, we used $\rho_{trad-uncoor}$ to measure the similarity between *traditional* and *uncoordinated* trends. *SRCC* assesses how well the relationship between two variables can be described using a monotonic function. A perfect *SRCC* of +1 (or -1) occurs when the variables are monotonically increasing (or decreasing) functions of the other. We measured $\rho_{trad-coor}$ and $\rho_{trad-uncoor}$ using three different q values (0.1, 0.3, 0.5), with all other variables fixed. Table 1 shows that as the social network exhibits an increasingly viral behavior with increasing q values, *structural* trends diverge from the traditional trends. The divergence is faster for *coordinated* than for *uncoordinated* trends.

Equation 4 evaluates the similarity the rankings of *all* the topics under two trend definitions. However, in most cases the rankings of unpopular topics is of little significance. Our goal is to identify top-k *coordinated* (*top–k_{coor}*) and *uncoordinated* (*top–k_{uncoor}*) topics. Therefore it is more important to observe the similarity between *top–k_{coor}* (or *top–k_{uncoor}*) and *top–k_{trad}*, i.e. *traditional* trends. In order to evaluate how good *top–k_{trad}* topics are at detecting *top–k_{coor}* (or *top–k_{uncoor}*), we use *average precision*, an IR technique used to evaluate score of a ranked list of documents for a query. *Average precision (AP)* incorporates precision and recall values while evaluating a top-k algorithm and can be computed as:

$$AP = \frac{\sum_{i=1}^{|D|} \text{Prec}(R_i)}{|D|} \tag{5}$$

Table 1: Model Similarity Statistics.

q	$\rho_{trad-coor}$	$\rho_{trad-uncoor}$	AP_{coor}	AP_{uncoor}
0.1	0.762	0.988	0.140	0.569
0.3	0.640	0.976	0.095	0.466
0.5	0.600	0.965	0.083	0.398

Table 2: Various Ranking Statistics.

		$AvgR_{trad}$	$AvgR_{coor}$	$AvgR_{uncoor}$
$p' = 0.1, q' = 0.1$	T'	24.44	36.52	18.56
$p'' = 0.032, q'' = 0.15$	T''	24.56	12.48	30.44
$p' = 0.1, q' = 0.1$	T'	24.64	12	34.68
$p'' = 0.2, q'' = 0.054$	T''	24.36	37	14.32

where $D = \{d_1, d_2, ..., d_k\}$ is the set of relevant documents, R is the ranked set of documents retrieved by the top-k algorithm and R_i is the set of ranked documents in R until document d_i is reached [25]. If d_i is not detected at all by the detection algorithm, $\text{Prec}(R_i) = 0$. We performed tests evaluating the *average precision* of $top\text{–}5_{trad}$ topics w.r.t. the relevant document set of $top\text{–}5_{coor}$ (or $top\text{–}5_{uncoor}$) topics. The results are given in Table 1 in columns AP_{coor} and AP_{uncoor} respectively and reflect similar results to these obtained using *SRCC* on the entire topic list. Similar experiments where all parameters except p values are fixed reveal that with increasing p values, similarity between *traditional* and *uncoordinated* trends increases. This adheres to the intuition that, as p values dominate q values, peer influence becomes less important and there is a smaller number of "spammy" topics for *uncoordinated* trends to filter out. For completeness purposes, the summary of the findings for this set of experiments is given in Appendix C.2.

Question 2: What is the nature of topics detected using structural trends? As we demonstrated earlier, structural trends tend to be different from traditional trends. But what do such differences corresponds to? In the previous experiments, the p and q values for each node and topic were set to the same value. Therefore, all the topics *had similar nature*. In the second set of experiments, half of the topics are set to one value ($q = p = 0.1$) while the other half of the topics are set to another. Consider a network G and a set of topics T. Let T' denote the set of topics in T with $q' = 0.1$ and $p' = 0.1$, and T'' denote the rest $T - T'$ topics. Setting $q < 0.1$ for a topic T_x results in T_x spreading less significantly through social ties and setting $p > 0.1$ balances this shortcoming by spreading through external influences. Therefore, for the set T'', there can be a distribution with $q'' < 0.1$ and $p'' > 0.1$ (or $q'' > 0.1$ and $p'' < 0.1$) values such that the average traditional ranking of T' and T'' are similar. Next we test, how topics in the subset T'' rank compared to T' w.r.t. their coordinated and uncoordinated scores. Let $AvgR_{trad}$, $AvgR_{coor}$ and $AvgR_{uncoor}$ of T' (or T'') denote the average ranking of topics that belong to T' (or T'') w.r.t. their f, g and h scores respectively. As the data set of this experiment consists of 50 topics, the topics rank from the highest score of 0 to 49. As could be expected, Table 2, shows that when $q'' < 0.1$ and $p'' > 0.1$, the *coordinated* significance of topics in set T'' is much lower than T', while *uncoordinated* significance is higher. The opposite behavior is observed when the settings are $q'' > 0.1$ and $p'' < 0.1$. For instance, an average of 12 among 25 topics in T'' indicate that all the topics in T'' rank in top-25 *coordinated* trends.

Possible use case of structural trends: detecting or filtering Sybil activity: Our experiments thus far modeled nodes of the network as having similar characteristics and focused on topics with varying properties. In reality, behavior of the nodes in the network can also vary. One use of structural trends would be if such variance in the form of spamming can be detected or filtered out using structural trends. Spam in social networks is an important and widely studied problem [7, 41, 16, 39, 27, 38]. We study one type of spamming behavior where a malicious user launches a Sybil attack [15] by creating a large number of virtual identities which have a large number of connections with other Sybil nodes and a small number of connections with real users. Given the importance of trends in social networks, it is important that the trending topics reported are not biased by the spam of a small number of Sybil nodes.

We claim that structural trends can identify or filter topics bolstered by malicious users in a Sybil setting. Unlike related work [38], our approach does not solely depend on the network graph and utilizes information diffusion data. Consider a social network where a group of Sybil nodes are trying to bolster the importance of a topic T_y. We performed experiments which investigate how the p and q values of Sybil nodes for topic T_y or the number of Sybil nodes effect the importance of T_y as a traditional, coordinated and uncoordinated trend. The details of the experimental study are given in Appendix C.1. Here we list some important findings: 1) The coordinated ranking of T_y is consistently higher than the traditional ranking which is higher than the uncoordinated ranking. This shows that *coordinated* trends identify Sybil activity while *uncoordinated* trends filter it. 2) Even with small values of p and q, there is a breakpoint at which T_y becomes drastically more popular as a coordinated trend. This breakpoint is observed at much larger p and q values for the other two definitions. 3) Similarly, the jump in the *coordinated* importance of T_y is observed with a small set of Sybil nodes whereas this jump is seen much later with the other definitions. These findings highlight the usefulness of structural trends in Sybil attack detection. In general, spamming behavior in social networks is not limited to Sybil attacks [27]. The effectiveness of structural trends under such different conditions is an open problem we plan to investigate in the future.

4.2 Analysis-Based Validation

Although using the methods introduced in Section 4.1, the value of structural trend definitions can be systematically studied, a verification using real data sets is crucial. We use a Twitter data set [31] of 467 million Twitter posts from 20 million users spanning a 7 month period containing author, time and content for each tweet. Using the Twitter social network graph published by Kwak et al. [22] we obtained the connections between the users sharing these tweets. We identified 2.7 million users that have at least one tweet that includes at least one hashtag. Such users have 230 million edges between them. We used hashtags to identify topics of tweets and observed that there were many hashtags that were similar except for some punctuation details or case differences. We categorize such hashtags under one topic. Although there are over 10 million hashtags in the Twitter data set, using this technique we were able to reduce this number to 2960495. We now summarize the key findings based on the analysis of this data.

Question 1: Are the *structural* trends different from *traditional* trends? Similar to the model-based verification, *SRCC* and *average precision* were computed to observe similarity of *structural* and *traditional* trends. The results are presented in Table 3. Consider the second column; ρ value for this case is *SRCC* of *traditional* and *coordinated* trends and $AP(29604)$, $AP(2960)$, and $AP(296)$ respectively represent the *average precision* of top 29604, 2960, 296 *traditional* topics in identifying the top 29604, 2960, 296 *coordinated* topics. These numbers correspond to the top-1, 0.1, 0.01 *percentile* of all the topics. The results consistently show that traditional trendiness is not a good predictor of *coordinated* trendiness, i.e. using *coordinated* trend definition, we identify topics that are not identified using *traditional* trend analysis. Likewise, the third column in Table 3 indicate that *uncoordinated* trends show significant differences from *traditional trends*. Interestingly, ad-

Table 3: Twitter Similarity Statistics.

	Coordinated	Uncoordinated
ρ	0.23	0.64
$AP(29604)$	0.54	0.9
$AP(2960)$	0.43	0.84
$AP(296)$	0.35	0.52

Table 4: Three Topics From Twitter.

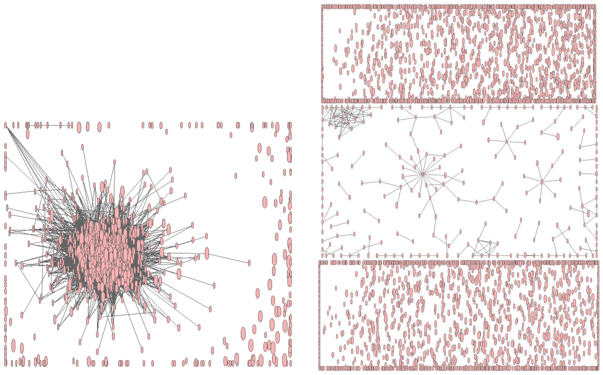
hashtag	R_{trad}	R_{coor}	R_{uncoor}	#users	#edges
#apple	78	201	2960491	24050	218451
#hhrs	82	10	20	5789	392501
#twitterafterdark	83	856	5	17441	37084

hering to the results obtained in Section 4.1, *uncoordinated trends* are more similar to *traditional trends* than *coordinated trends*.

Question 2: What is the nature of topics detected using structural trends? As our goal is to detect interesting topics using *structural* trends that would be undetected otherwise, we studied a set of topics in Twitter data set that have sizable number of mentions though not ranking high enough to be detected as a *traditional* trend (ranking 60^{th} to 100^{th}). Of those topics we identified topics that have a high structural significance compared to their *traditional* significance, i.e., top-10 topics sorted by $R_{coor}(x) - R_{trad}(x)$ (or $R_{uncoor}(x) - R_{trad}(x)$) where $R_{trad}(x)$, $R_{coor}(x)$ and $R_{uncoor}(x)$ corresponds to the *traditional, coordinated* and *uncoordinated* ranking of a topic T_x respectively. The results indicate that *coordinated* trends result from tweets from a relatively small number of users (7694 on average) with a very large number of ties (21.5 number of neighbors per node on average), whereas *uncoordinated trends* result from tweets from a large number of distinct users (21114 on average) with small number of ties (8.6 number of neighbors per node on average). The details of such topics are provided in Appendix C.2 and show interesting distinctions.

Table 4 demonstrates the results using three topics; #apple, #twitterafterdark, #hhrs. The columns of the table correspond to the hashtag, traditional, coordinated and uncoordinated ranking of the hashtag, number of distinct users that used the hashtag and the number of edges between such users respectively. Of those three topics that have similar traditional scores, #twitterafterdark has a high uncoordinated score, #hhrs has a high coordinated score and #apple is insignificant as a structural trend. The *coordinated* trend #hhrs originates from a small number of nodes with a large number of edges between them whereas the *uncoordinated* trend #twitterafterdark originates from a large number of distinct users with a much smaller ratio of edges between them. The hashtag #hhrs refers to Hugh Hewitt Radio Show which is a conservative talk show, whereas #twitterafterdark is a hashtag used by users to refer to their experiences at night. It is intuitive that #hhrs is mentioned by connected pairs of nodes, especially considering the effects of homophily in connection formation [23]. On the other hand, #twitterafterdark is an idiom whose usage depends on experiences of users rather than the use of the hashtag by their friends. The hashtag #apple is used in tweets relating to Apple products. These three hashtags have very different characteristics that would go unnoticed with traditional trend analysis as their traditional rank is similar.

We now give a visualization that demonstrates the difference between trends detected using *coordinated* and *traditional* trends. We use Prefuse, an open-source software [18], to visualize subgraphs of the Twitter data set, consisting of only nodes that participated in particular hashtags and edges between such nodes. The size of a node is proportional to $\log_2$ of number of tweets that node had on that particular hashtag. The visualization results for hashtags #pawpawty and #mafiawars are given in Figures 2(a) and 2(b). These two hashtags have different categorical natures; #pawpawty is commonly used to raise money for animal rescue organizations, whereas #mafiawars is commonly used by gamers. Despite this important difference, the two hashtags have very similar traditional rankings of 289 and 212 respectively. Unlike *traditional* trends, *coordinated* trends detect the difference between the two hashtags; #pawpawty has a high *coordinated* importance, ranking 24^{th}, while #mafiawars does not. This difference can be seen in Figures 2(a) and 2(b). We can see that #mafiawars has a large number of unconnected nodes, while the opposite is true for #pawpawty suggesting that this socially motivated hashtag, unlike #mafiawars, diffuses mostly through the friendship edges or that homophily effect for #pawpawty is stronger. This analysis takes us to our next question: Do hashtags with different categorical characteristics have lower or higher structural importance?

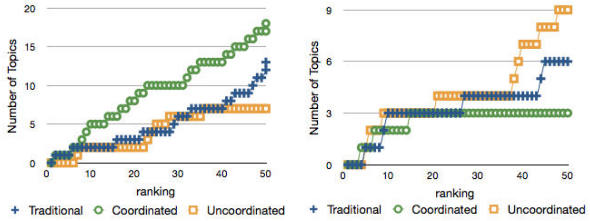


(a) #pawpawty

(b) #mafiawars

Figure 2: Two traditionally similar hashtags in Twitter.

In order to answer this question, we analyze 500 hashtags that are categorized into 7 different topics; political, technology, celebrity, games, idioms, movies, music and none. These hashtags and their categories were obtained from a recent study [28] that categorizes the top 500 hashtags for a Twitter data set. The timeframe of the data set used in that study overlaps with our data set. Therefore these hashtags have significant importance in our data set as well, though not necessarily amongst top-500. Our analysis provides some interesting insights as to how people use Twitter to share information. Figure 3(a) demonstrates the cumulative distribution function (CDF) of coordinated, uncoordinated and traditional ranking of political hashtags given in [28]. We see that using the coordinated trends definition, the importance of political hashtags are bolstered. Political hashtags having a high coordinated importance indicate that such hashtags are used by highly clustered set of nodes and that people tend to re-share political information shared by their friends. This is not the case for other categories, such as idioms [28] as demonstrated in Figure 3(b). Such topics are more significant as an *uncoordinated* trend indicating that usage of these hashtags is mostly amongst dispersed nodes. This phenomenon could have been facilitated by the unique nature of Twitter, which broadcasts tweets of every user. A user interested in a specific hashtag can easily obtain the list of tweets under that hashtag. So usage of Twitter is likely to be more centric to the use of this feature rather than social friend following. A different behavior can possibly be observed in other social networks. In general, the structural trend based data analysis facilitates the identification of interesting specifics of user interactions in different categorical contexts in various social network. An interesting problem is the evolution of trends throughout time which can be studied using sliding windows of user tweets. More interestingly online solutions in this setting can be applied using techniques similar to the ones explained in Section 5. Due to space limitations, we omit such analyses and leave an extensive study on timing issues as future work.



(a) Political hashtags (b) Idiom hashtags

Figure 3: CDF of ranking of topics of different categories.

5. STRUCTURAL TREND DETECTION

In this section, we provide methods for both *coordinated* and *uncoordinated* trend detection. In Section 5.1, we will first give details about the solution for *coordinated* trends. Later in Section 5.2, we will provide the details for *uncoordinated* trend detection.

5.1 Coordinated Trend Detection

We start by presenting a naive solution to detect *coordinated trends*, i.e. computing Equation 2 exactly for each topic. Since this solution is expensive for large social networks with high traffic of information sharing, we next explore ways to improve efficiency. To this end, we propose a sampling based solution. We show that a simple sampling method can be used while still guaranteeing high accuracy, especially for *popular* topics. In order to demonstrate the use of this sampling technique, we reduce the problem of evaluating the *coordinated trendiness* of each topic to the problem of counting local triangles, i.e., counting the number of triangles incident to a given node in a graph G.

5.1.1 Incremental Counting Algorithm

As our main goal is to detect trends, it is crucial to provide incremental solutions. Therefore semi-streaming methods [34, 6] which traverse the data a non-constant number of times are not applicable. For these methods, updates such as the receipt of a small number of broadcasts would necessitate the repetition of the whole process to find new trends. Instead, we propose using an incremental approach. The approach introduced in this section finds *exact values* and therefore is computationally expensive, but using the sampling method described in Section 5.1.2, the complexity can be reduced.

Consider the actions that need to be taken upon receiving a new tuple $\langle n_l, T_x\rangle$. Assume that until this point the *exact* value of $C_{i,x}$ for each $T_x \in T$ and $n_i \in N$ and the *exact* value for Equation 2 for each T_x are known. Upon the receipt of $\langle n_l, T_x\rangle$, $C_{l,x}$ has to be incremented by 1. Also, the score of T_x should be updated as:

$$g'(T_x) = g(T_x) + \sum_{n_i \in N_l'} C_{i,x} + \sum_{n_i \in N_l} C_{i,x} \qquad (6)$$

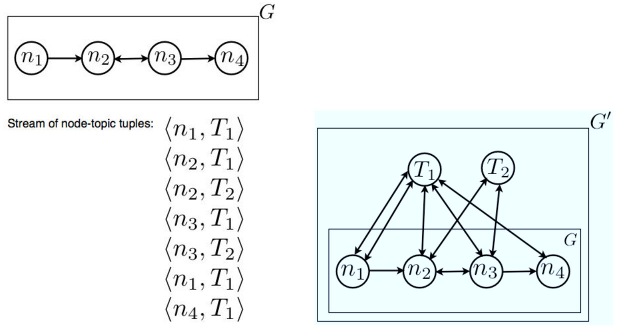
where $g(T_x)$ is the *coordinated* score of T_x before receipt of $\langle n_l, T_x\rangle$, $g'(T_x)$ is its score afterwards, $N_i = \{n_j | e_{i,j} \in E\}$ and $N_i{}' = \{n_j | e_{j,i} \in E\}$. The proof of the correctness of this equation can be found in Section B.1. As is evident from this computation, after receiving tuple $\langle n_l, T_x\rangle$, $g(T_x)$ has to be increased by the sum of all $C_{j,x}$ where n_j is a neighbor of n_l, and $C_{m,x}$, where n_l is a neighbor of n_m. This requires O(n) reads. However, in social networks, only a small fraction of nodes are connected to a large number of nodes. So in most cases this operation requires a small number of reads. The solution requires using two adjacency lists per node n_i, one to keep track of incoming and outgoing edges. Fast access to $C_{i,x}$ for each i and x is needed as well. Therefore a hashtable per topic is used to keep track of the counts of broadcasts per node.

As our ultimate goal is to give an ordered list of *top-k* coordinated topics *at each point in time*, in addition to accurately reporting coordinated scores per topic we need to provide a sorted representation of *top-k* topics. Therefore, a simple solution uses a sorted structure to keep track of the *top-k* topics. In this case the receipt of a new tuple $\langle n_l, T_x\rangle$ might require an update to this structure as well. The naive implementation provides a good solution for small networks with a small number of broadcasts per seconds. However, the sheer volume of information shared on online social networks today still poses a scalability challenge. A recent report from Twitter announced 3283 tweets per second [33]. Data flow at this scale calls for solutions that sacrifice accuracy for efficiency.

5.1.2 Counting Local Triangles and Sampling

We now propose our sampling based solution to overcome the scalability challenge of coordinated trend detection. As it is easier to describe the correctness in a graph-oriented manner, we will show that the problem of finding *coordinated trends* is equivalent to counting local triangles in a multi-graph. Later, we will prove that using sampling this specific problem can be made more efficient.

Consider a directed graph $G = (N, E)$, a set of topics T and stream of tuples S, where a tuple is in the form: $\langle n_i, T_x\rangle$ s.t. $n_i \in N$ and $T_x \in T$. Create a directed multi-graph $G' = (N', E')$ s.t. $N' = N \cup T$ and $E' = \{(u,v) | (u,v) \in E \vee \langle u,v\rangle \in S \vee \langle v,u\rangle \in S\}$. The nodes can be categorized into two categories: *topic nodes* $T_x \in T$ and *user nodes* $n_i \in N$. Let the edges of the form (n_i, T_x) (or (T_x, n_i)) be *topic edges* and denote this set as E_t. Similarly, edges of the form (n_i, n_j) are *friendship edges* and are denoted by E_f. Clearly $E' = E_t \cup E_f$. In a multi-graph two vertices may be connected by more than one edge. By construction of G', there can be at most one *friendship* edge from a node n_i to n_j and an arbitrary number of *topic* edges between n_i and T_x. Any three nodes u, v and w s.t. $(u,v) \in E' \wedge (v,w) \in E' \wedge (w,u) \in E'$ form a *triangle* in G'. The $g(T_x)$ score of a topic T_x given in Equation 2 is simply the number of triangles incident to node T_x in G'. Figure 4 gives an example of one such reduction. Nodes T_1, n_2, n_3 induce two triangles whereas T_1, n_3, n_4 induce only one since (n_2, n_3) is a bidirectional edge whereas (n_3, n_4) is unidirectional. Also T_1, n_1, n_2 induce two triangles because there are two *topic* edges between T_1 and n_1.


G
n_1
n_2
n_3
n_4
Stream of node-topic tuples:
$\langle n_1, T_1\rangle$
$\langle n_2, T_1\rangle$
$\langle n_2, T_2\rangle$
$\langle n_3, T_1\rangle$
$\langle n_3, T_2\rangle$
$\langle n_1, T_1\rangle$
$\langle n_4, T_1\rangle$
G'
T_1
T_2
G
n_1
n_2
n_3
n_4

(a) A graph G and stream of mentions S.

(b) The multigraph created using G and S.

Figure 4: Reduction to counting local triangles.

After this reduction, the stream of $\langle n_i, T_x\rangle$ tuples can be observed as the incoming *topic* edges of G'. Given the entire graph G is available and only E_t, the *topic edges* are sampled, $g(T_x)$ can be accurately predicted. The procedure is straightforward and the sampling method resembles to the one introduced in [35]: Create a directed multi-graph $G'' = (N'', E'')$ s.t. $N'' = N'$ and $E'' = \{(u,v) | (u,v) \in E\}$. For each incoming tuple $\langle n_i, T_x \rangle$, which corresponds to *topic* edges (n_i, T_x) and (T_x, n_i) in G', flip a coin with bias p_s. With p_s probability, we keep *both* (n_i, T_x) and (T_x, n_i) edges by setting $E'' = E'' \cup (n_i, T_x) \cup (T_x, n_i)$ and discard them both otherwise. The number of triangles involving T_x in G' can be estimated as $X_x = Count_x / {p_s}^2$, where $Count_x$ denotes the number of triangles involving T_x in G''. We can guarantee that the number of triangles calculated based on the sampled data is a good approximation of the actual number of triangles. Specifically, the probability that the prediction X_x is off by $\varepsilon\Delta_x$ is upper-bounded by the following equation:

$$\Pr(|X_x - \Delta_x| \geq \varepsilon\Delta_x) \leq \frac{\mathrm{Var}(X_x)}{\varepsilon^2 {\Delta_x}^2} \leq \frac{({p_s}^2 - {p_s}^4)}{{p_s}^4 \varepsilon^2 \Delta_x} + 2\alpha_x \frac{({p_s}^3 - {p_s}^4)}{{p_s}^4 \varepsilon^2 {\Delta_x}^2} \tag{7}$$

where Δ_x is the actual number of triangles involving T_x, α_x is the number of pairs of triangles that involve T_x and are not edge disjoint and p_s is the rate of sampling. The proof of correctness of Equation 7 is provided in Appendix A. As is evident from Equation 7 the quality of the estimate depends on the number of triangles as well as the number of edge-disjoint triangles. Since the number of multi-edges has a big effect on this property, the quality of the estimate depends on the number of times a specific user mentions a specific topic. As this number gets increasingly large, the quality of the estimate degrades. However the estimate becomes quadratically better with increasing Δ_x and only linearly worse with α_x which is smaller so the estimate is still better for "trendy" topics.

5.2 Uncoordinated Trend Detection

Similar to *coordinated* trends, *uncoordinated* trends can be reduced to counting local triangles in a multi-graph. Considering same definition for G, T and S as given in Section 5.1, create a multi-graph $G' = (N', E')$ s.t. $N' = N \cup T$ and $E' = \{(u,v) | (u,v) \notin E \vee (u,v) \in S\}$. The $h(T_x)$ is simply the number of triangles incident to node T_x in G'. In this setting, the tuples $\langle n_i, T_x \rangle$ are the edges of the multi-graph G'. As demonstrated in Section 5.1.2, this problem can be efficiently approximated by sampling. Since an online algorithm is a requirement, the *uncoordinated* trendiness score of topics should be incrementally updated. As proven in Section B.2, the increase can be calculated in the following way:

$$h'(T_x) = h(T_x) + \sum_{n_i \in N \setminus (n_l \cup N_l)} C_{i,x} + \sum_{n_i \in N \setminus (n_l \cup N_l{}')} C_{i,x} \tag{8}$$

where $h(T_x)$ is the *uncoordinated* score of T_x before receipt of $\langle n_l, T_x \rangle$, $h'(T_x)$ is its score after the receipt of the tuple, $N_i = \{n_j | e_{i,j} \in E\}$and $N_i{}' = \{n_j | e_{j,i} \in E\}$. Upon receiving tuple $\langle n_l, T_x \rangle$, the *uncoordinated trendiness* score of T_x has to be increased by the sum of all $C_{j,x}$ such that n_j is not a neighbor of n_l and $C_{m,x}$ such that n_l is not a neighbor of n_m. This requires in the worst case $O(n)$ reads. Unfortunately unlike the computation necessary *coordinated trendiness*, this operation in most cases requires close to n reads. However, a simple realization results in an efficient solution that performs a small number of reads per update by making use of the power-law degree distribution of social networks. By keeping track of *traditional trendiness* score, $f(T_x)$, for each topic T_x, the update on $h(T_x)$ can be computed as: $2 * f(T_x) - \sum_{n_j \in N_l} C_{j,x} - \sum_{n_j \in N_l{}'} C_{j,x} - 2 * C_{l,x}$ as shown in Section B.2.

5.3 Experimental Results on Twitter

Our goal is to provide a ranked list of top-k *structural* trends. Therefore, we performed experiments on the Twitter data set introduced in Section 4.2 to compute the *average precision (AP)* of sampled data for both *coordinated* and *uncoordinated* top-k lists for different values of p_s (0.5,0,2,0.1, 0.01 and 0.005) and k. Figure 5, which provides the results for *coordinated* trends, shows that top-29 *coordinated* trend detection is largely robust to the sampling parameter, i.e. even for a small value of $p_s = 0.005$ where approximately 1 out of 200 tuples is processed, *AP* lies above 0.93. This is not the case for the top-29604 topics where *AP* degrades largely with decreasing p_s. This is mostly due to the large number of *tail* topics that are unpopular and have close-to-zero values. This behavior therefore is to be expected considering that sampling has low accuracy for unpopular topics as is shown in Equation 7. Note, however, unpopular topics are of little interest for trend detection. Results for *uncoordinated* trends are similar to that of *coordinated* trends and are provided in Appendix C.3. Interestingly, *uncoordinated* trend detection is more robust to sampling since the quality of sampling is higher for larger values of exact number of triangles as given in Equation 7 and due to the sparsity of social network graphs number of triangles induced from *uncoordinated* trendiness tend to be larger than that of *coordinated* trendiness. As could be expected, a linear speed-up is observed w.r.t $1/p_s$. We refer the reader to Appendix C.3 for the experiment results.

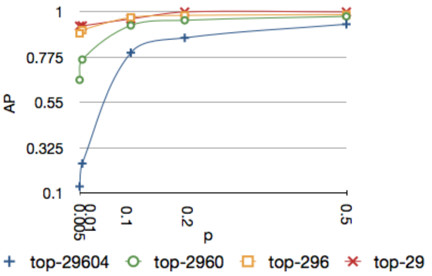


Figure 5: Average Precision of sampling for coordinated trends.

6. CONCLUSION

In this paper, we introduced new methods for identification of important topics in social networks that utilizes the network topology. We proposed two novel trend definitions called *coordinated* and *uncoordinated* trends that detect topics that are popular amongst highly clustered and distributed users respectively. We also introduced a novel information diffusion model called *Independent Trend Formation Model (ITFM)* that distinguishes viral diffusion of information from diffusion through external entities such as news media and captures the diffusion of an arbitrary number of topics in a social network. Using both *ITFM* and a Twitter data set with 41.7 million nodes and 417 million posts, we demonstrated the value of the new trend definitions by showing that they identify substantially different set of topics as trending and shed light on the way information diffuses in social networks. We also proposed a sampling technique for structural trend detection that provides computational gain as well as a solution within an acceptable error bound. Our experimental study on the Twitter data set show an impressive 0.93 *average precision* for *coordinated* trends and a perfect *average precision* of 1 for *uncoordinated* trends.

As this work emphasizes, traditional ways of identifying popular items are not enough to characterize information diffusion and discover interesting activity in social networks. We specifically explored two novel ways of discovering trends. We do not propose the new techniques as a substitute for traditional trend detection but rather as a compliment. As future work, we plan to study more general structural trend definitions that explore the space between the two extremes introduced in this work, such as topics discussed by a group of *c* connected users. We also plan to investigate other metrics such as strength of ties and timing of mentions to incorporate into structural trend analysis.

7. REFERENCES

[1] R. Agrawal, M. Potamias, and E. Terzi. Learning the nature of information in social networks. Technical Report MSR-TR-2011-59, Microsoft Research, May 2011.

[2] J. Allan, editor. *Topic detection and tracking: event-based information organization*. Kluwer Academic Publishers, Norwell, MA, USA, 2002.

[3] N. Alon and J. Spencer. *The probabilistic method*. Wiley-Interscience, 2000.

[4] N. Alon, R. Yuster, and U. Zwick. Finding and counting given length cycles. *Algorithmica*, 17:209–223, 1997.

[5] Z. Bar-Yossef, R. Kumar, and D. Sivakumar. Reductions in streaming algorithms, with an application to counting triangles in graphs. In *SODA '02*, pages 623–632, 2002.

[6] L. Becchetti, P. Boldi, C. Castillo, and A. Gionis. Efficient semi-streaming algorithms for local triangle counting in massive graphs. In *KDD '08*, pages 16–24, 2008.

[7] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. Detecting spammers on twitter. In *CEAS*, 2010.

[8] S. Bharathi, D. Kempe, and M. Salek. Competitive influence maximization in social networks. In *WINE*, pages 306–311, 2007.

[9] S. P. Borgatti. Centrality and network flow. *Social Networks*, 27(1):55 – 71, 2005.

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

[11] C. Budak, D. Agrawal, and A. El Abbadi. Limiting the spread of misinformation in social networks. In *WWW '11*, pages 665–674, 2011.

[12] L. S. Buriol, G. Frahling, S. Leonardi, A. Marchetti-Spaccamela, and C. Sohler. Counting triangles in data streams. In *PODS '06*, pages 253–262, 2006.

[13] T. Carnes, C. Nagarajan, S. M. Wild, and A. van Zuylen. Maximizing influence in a competitive social network: a follower's perspective. In *ICEC '07*, pages 351–360, 2007.

[14] G. Cormode and M. Hadjieleftheriou. Finding the frequent items in streams of data. *Commun. ACM*, 52:97–105, October 2009.

[15] J. R. Douceur. The sybil attack. In *IPTPS*, pages 251–260, 2002.

[16] C. Grier, K. Thomas, V. Paxson, and M. Zhang. @spam: the underground on 140 characters or less. In *ACM CCS*, pages 27–37, 2010.

[17] S. Havre, B. Hetzler, and L. Nowell. ThemeRiver: visualizing theme changes over time. In *InfoVis 2000*, pages 115–123, 2000.

[18] J. Heer, S. Card, and J. Landay. Prefuse: a toolkit for interactive information visualization. In *CHI'05*, pages 421–430, 2005.

[19] Tweetstats. `http://tweetstats.com/trends`.

[20] H. Jowhari and M. Ghodsi. New streaming algorithms for counting triangles in graphs. In *COCOON'05*, pages 710–716, 2005.

[21] D. Kempe, J. M. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *KDD'03*, pages 137–146, 2003.

[22] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *WWW '10*, pages 591–600, 2010.

[23] T. La Fond and J. Neville. Randomization tests for distinguishing social influence and homophily effects. In *WWW'10*, pages 601–610, 2010.

[24] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *KDD '09*, pages 497–506, 2009.

[25] C. D. Manning, P. Raghavan, and H. Schtze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[26] J. Maritz. *Distribution-free statistical methods*. Chapman & Hall/CRC, 1995.

[27] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, S. Patil, A. Flammini, and F. Menczer. Detecting and tracking the spread of astroturf memes in microblog streams. *CoRR*, abs/1011.3768, 2010.

[28] D. M. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *WWW '11*, pages 695–704, 2011.

[29] A. Sala, L. Cao, C. Wilson, R. Zablit, H. Zheng, and B. Y. Zhao. Measurement-calibrated graph models for social network experiments. In *WWW '10*, pages 861–870, 2010.

[30] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. Twitterstand: news in tweets. In *GIS '09*, pages 42–51, 2009.

[31] Snap: Network datasets: 476 million twitter tweets. `http://snap.stanford.edu/data/twitter7.html`.

[32] B. E. Teitler, M. D. Lieberman, D. Panozzo, J. Sankaranarayanan, H. Samet, and J. Sperling. Newsstand: a new view on news. In *GIS '08*, pages 1–10, 2008.

[33] Another big record: Part deux. `http://blog.twitter.com/2010/06/another-big-record-part-deux.html`.

[34] C. E. Tsourakakis. Fast counting of triangles in large real networks without counting: Algorithms and laws. In *ICDM '08*, pages 608–617, 2008.

[35] C. E. Tsourakakis, U. Kang, G. L. Miller, and C. Faloutsos. Doulion: counting triangles in massive graphs with a coin. In *KDD '09*, pages 837–846, 2009.

[36] Twitter. `http://twitter.com/`.

[37] Why twitter hashtags and trending topics are useless to marketers. `http://blog.hubspot.com/blog/tabid/6307/bid/4694/Why-Twitter-Hashtags-and-Trending-Topics-Are-Useless-to-Marketers.aspx`.

[38] B. Viswanath, A. Post, K. P. Gummadi, and A. Mislove. An analysis of social network-based sybil defenses. *SIGCOMM Comput. Commun. Rev.*, 40:363–374, August 2010.

[39] A. H. Wang. Don't follow me - spam detection in twitter. In *SECRYPT*, pages 142–151, 2010.

[40] Twitter: We are not keeping wikileaks out of trending topics. `http://mashable.com/2010/12/06/wikileaks-twitter-censorship/`.

[41] S. Yardi, D. Romero, G. Schoenebeck, and D. Boyd. Detecting spam in a twitter network. *First Monday*, 15(1):1–13, 2010.

[42] H. Yu, M. Kaminsky, P. B. Gibbons, and A. Flaxman. Sybilguard: defending against sybil attacks via social networks. *SIGCOMM Comput. Commun. Rev.*, 36:267–278, 2006.

Table 5: Definitions of symbols and Acronyms.

Symbol	Definition
$G = (N, E)$	social network graph
S	stream of node-topic tuples (mentions)
T	Topic nodes in G' (induced from the set of all possible topics)
$G' = (N', E')$	multi-graph induced using G, T and S
N'	$= N \cup T$
E'	$= E_f \cup E_t$ where E_f is the set of friendship edges and E_t is the set of topic edges
G''	multi-graph after sampling
Δ_x	number of triangles involving T_x in $G\prime$
$\delta_{x,j}$	indicator variable, $\delta_{x,j} = 1$ if j^{th} triangle of T_x exists in G'' and $\delta_{x,j} = 0$ otherwise
X_x	estimate of number of triangles after sampling (computed as $Count_x/p^2$, where $Count_x$ is the number of triangles involving T_x in G'')
p_s	sampling probability, i.e. probability that a topic edge in G' exists in G''

APPENDIX

A. PROOF OF QUALITY OF SAMPLING

In this section we prove the correctness of Equation 7 which shows the error bound for counting the number of local triangles in G' where a certain subset of the edges (E_t) are sampled. We list an overview of notations in Table 5 as a guideline for this section and refer the reader to Section 5.1.2 for the construction of G' and G''. Our goal is to prove that the number of local triangles involving a specific *topic node* T_x can be accurately estimated given that the *topic edges* in E_t are sampled and the exact sets of N' and E_f are provided. To this end, we start by studying the mean and variance of the number of triangles estimated using the sampled data and derive bounds on the expected number of triangles detected.

THEOREM A.1. *The expected value of X_x in G'' is Δ_x which is equivalent to the score of topic T_x.*

PROOF. X_x is the multiplication of the sum of indicator variables for topic T_x and $(1/p_s)^2$. Therefore, $\mathrm{E}[X_x] = \mathrm{E}[\sum_{j=0}^{\Delta_x} \delta_{x,j}/{p_s}^2]$ $= \sum_{j=0}^{\Delta_x} \mathrm{E}[\delta_{x,j}/{p_s}^2] = 1/{p_s}^2 \sum_{j=0}^{\Delta_x} \mathrm{E}[\delta_{x,j}] = 1/{p_s}^2 \sum_{j=0}^{\Delta_x} {p_s}^2 = \Delta_x$ □

Using Chebyshev's inequality [3], which states $\Pr(|X_x - \Delta_x| \geq \varepsilon\Delta_x) \leq \frac{\mathrm{Var}(X_x)}{\varepsilon^2 {\Delta_x}^2}$, guarantees of the accuracy of X_x in predicting the actual Δ_x values can be given. In order to do so, we now study the variance of variable X_x.

THEOREM A.2. *The variance of X_x is:*

$$\mathrm{Var}(X_x) = \frac{\Delta_x({p_s}^2 - {p_s}^4) + 2\alpha_x({p_s}^3 - {p_s}^4)}{{p_s}^4}$$

where α_x is the number of pairs of triangles that involve T_x and are not edge disjoint.

PROOF. X_x is a sum of indicators that a certain triangle involving T_x survives after sampling. These indicators are not independently distributed. Consider two triangles denoted by indicator variables $\delta_{x,j}$ and $\delta_{x,l}$ where $j \neq l$ (j^{th} and l^{th} triangle involving T_x). Such two triangles cannot share all three edges as they are distinct triangles. They can neither share two *friendship edges* since there can be at most 1 edge from a node n_i to n_j. They cannot share two *topic edges* either since in this case the triangles would be identical as two triangles sharing two topic edges would also have to share the friendship edge. Eliminating such possibilities, there are four possible cases to be considered. They can share: 1) one *topic edge* ($\delta_{i,1}$ and $\delta_{i,2}$ in Figure 6), 2) one *friendship edge* 3) one *friendship* and one *topic edge* or 4) no edge ($\delta_{i,3}$ and $\delta_{i,4}$ in Figure 6). Figure 6 lists these possible scenarios of how two such indicators might (or not) be dependent. For case 2) and 4), the two indicators would be independent as friendship edges are not sampled. For cases 1) and 3), the two indicator variables both are dependent on the *topic edge* "surviving". Let number of cases of the form 1) or 3) be α_x for topic T_x. The variance of X can be computed as:

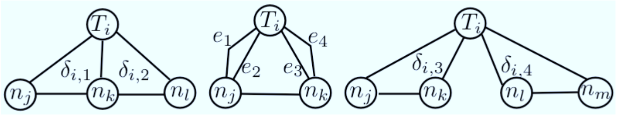


Figure 6: Cases to be considered for variance.

$$\mathrm{Var}(X_x) = \mathrm{Var}(\frac{1}{p^2}\sum_{j=1}^{\Delta_x} \delta_{x,j}) = \frac{1}{p^4}\sum_{j=1}^{\Delta_x}\sum_{l=1}^{\Delta_x} Cov(\delta_{x,j}, \delta_{x,l})$$

There are ${\Delta_x}^2$ terms in this summation. Δ_x of these terms are the variances of indicator variables. Since there are α_x of cases where two indicator variables are dependent on each other (share one *topic edge*), the covariance for α_x out of $\binom{\Delta_x}{2}$ pairs of indicator variables is: $\mathrm{Cov}(\delta_{x,j}, \delta_{x,l}) = {p_s}^3 - {p_s}^4$. $\mathrm{Cov}(\delta_{x,m}, \delta_{x,o}) = {p_s}^4 - {p_s}^4 = 0$ for the rest $\binom{\Delta_x}{2} - \alpha_x$ terms. Therefore the variance is:

$$\mathrm{Var}(X_x) = \frac{1}{{p_s}^4}(\Delta_x({p_s}^2 - {p_s}^4) + 2\alpha_x({p_s}^3 - {p_s}^4))$$

□

Using Chebyshev's inequality [3] and Theorems A.2 and A.1, we can show the correctness of Equation 7 as follows:

$$\Pr(|X_x - \Delta_x| \geq \varepsilon\Delta_x) \leq \frac{\mathrm{Var}(X_x)}{\varepsilon^2 {\Delta_x}^2} \leq \frac{({p_s}^2 - {p_s}^4)}{{p_s}^4\varepsilon^2\Delta_x} + 2\alpha_x\frac{({p_s}^3 - {p_s}^4)}{{p_s}^4\varepsilon^2{\Delta_x}^2}$$

B. INCREMENTAL COORDINATED AND UNCOORDINATED SCORE UPDATES

In this section we prove the correctness of Equations 6 and 8 which identify how *coordinated* and *uncoordinated* scores of a topic T_x need to be update upon receipt of a tuple $\langle n_l, T_x \rangle$.

B.1 Coordinated Score Update

Upon processing the *mention* $\langle n_l, T_x \rangle$, the "trendiness score" of T_x has to be increased by the sum of all $C_{j,x}$ such that n_j is a neighbor of n_l and $C_{m,x}$ such that n_l is a neighbor of n_m. Now, we will prove the correctness of this update function which is given in Equation 6 where $g(T_x)$ denotes the *coordinated* score of T_x before receipt of $\langle n_l, T_x \rangle$ and $g'(T_x)$ denotes its score afterwards. Similarly, we use the symbol C for the counts of mentions per node before receipt of the new mention and C' for afterwards. Keeping in mind that the only C' value changed is $(C_{l,x})'$, the new *coordinated* score for T_x can be calculated in the following way:

$$\begin{aligned} g'(T_x) &= \sum_{\substack{n_i \in N \\ n_j \in N_i}} C_{i,x}' \cdot C_{j,x}' \\ &= \sum_{\substack{n_i \in N \backslash n_l \\ n_j \in N_i \backslash n_l}} C_{i,x}' \cdot C_{j,x}' + \sum_{\substack{n_i = n_l \\ n_j \in N_i}} C_{i,x}' \cdot C_{j,x}' + \sum_{\substack{n_j = n_l \\ n_i \in N_j'}} C_{i,x}' \cdot C_{j,x}' \\ &= \sum_{\substack{n_i \in N \backslash n_l \\ n_j \in N_i \backslash n_l}} C_{i,x} \cdot C_{j,x} + \sum_{n_j \in N_l} (C_{l,x} + 1) \cdot C_{j,x} + \sum_{n_i \in N_l'} C_{i,x} \cdot (C_{l,x} + 1) \\ &= \sum_{\substack{n_i \in N \\ n_j \in N_i}} C_{i,x} \cdot C_{j,x} + \sum_{n_i \in N_l'} C_{i,x} + \sum_{n_i \in N_l} C_{i,x} \\ &= g(T_x) + \sum_{n_i \in N_l'} C_{i,x} + \sum_{n_i \in N_l} C_{i,x} \end{aligned}$$

where $N_i = \{n_j | e_{i,j} \in E\}$ and $N_i' = \{n_j | e_{j,i} \in E\}$.

B.2 Uncoordinated Score Update

According to Equation 8, upon receiving mention $\langle n_l, T_x \rangle$, the *uncoordinated trendiness* score of T_x has to be increased by the sum of all $C_{j,x}$ such that n_j is not a neighbor of n_l and $C_{m,x}$ such that n_l is not a neighbor of n_m. Now we will prove this statement. Similar to Section B.1, the counts per node are denoted by C before receipt of the new tuple and C' after receipt of the new tuple.

$$\begin{aligned} h'(T_x) &= \sum_{\substack{n_i \in N \\ n_j \in N_i^c}} C_{i,x}' \cdot C_{j,x}' \\ &= \sum_{\substack{n_i \in (N \backslash n_l) \\ n_j \in (N_i^c \backslash n_l)}} C_{i,x}' \cdot C_{j,x}' + \sum_{\substack{n_i = n_l \\ n_j \in N_i^c}} C_{i,x}' \cdot C_{j,x}' + \sum_{\substack{n_j = n_l \\ n_i \in N_j'^c}} C_{i,x}' \cdot C_{j,x}' \\ &= \sum_{\substack{n_i \in (N \backslash n_l) \\ n_j \in (N_i^c \backslash n_l)}} C_{i,x} \cdot C_{j,x} + \sum_{n_j \in N_i^c} (C_{l,x} + 1) \cdot C_{j,x} + \sum_{n_i \in N_l'^c} C_{i,x} \cdot (C_{l,x} + 1) \\ &= \sum_{\substack{n_i \in N \\ n_j \in N_i^c}} C_{i,x} \cdot C_{j,x} + \sum_{n_i \in N \backslash (n_l \cup N_l)} C_{i,x} + \sum_{n_i \in N \backslash (n_l \cup N_l')} C_{i,x} \\ &= h(T_x) + \sum_{n_i \in N \backslash (n_l \cup N_l)} C_{i,x} + \sum_{n_i \in N \backslash (n_l \cup N_l')} C_{i,x} \end{aligned}$$

where $h(T_x)$ is the *uncoordinated* score of T_x before receipt of the new tuple $\langle n_l, T_x \rangle$, $h'(T_x)$ is its score after the receipt of the tuple, $N_i = \{n_j | e_{i,j} \in E\}$, $N_i' = \{n_j | e_{j,i} \in E\}$, $N_i^c = N \setminus (n_i \cup N_i)$ and $N_i'^c = N \setminus (n_i \cup N_i')$. We can further analyze this result to obtain:

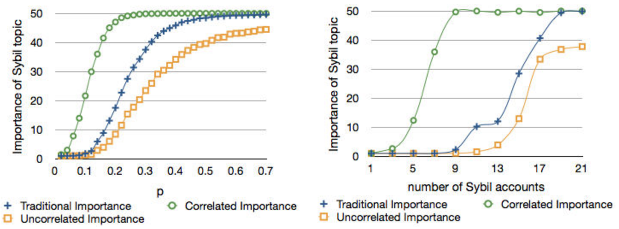
$$\begin{aligned} h'(T_x) &= h(T_x) + \sum_{n_i \in N \backslash (n_l \cup N_l)} C_{i,x} + \sum_{n_i \in N \backslash (n_l \cup N_l')} C_{i,x} \\ &= h(T_x) + \sum_{n_i \in N} C_{i,x} - C_{l,x} - \sum_{n_i \in N_l} C_{i,x} + \sum_{n_i \in N} C_{i,x} - C_{l,x} - \sum_{n_i \in N_l'} C_{i,x} \\ &= h(T_x) + 2 * f(T_x) - \sum_{n_i \in N_l} C_{i,x} - \sum_{n_i \in N_l'} C_{i,x} - 2 * C_{l,x} \end{aligned}$$

Using this observation, the power-law properties of degree distribution of social networks can be leveraged to perform the update on $h(T_x)$ requiring only a small number of reads for most cases.

C. FURTHER EXPERIMENTAL RESULTS AND ANALYSES

C.1 Sybil Attack Experiments

In Section 4, we claimed that structural trend analysis has a nice side effect of either identifying or filtering topics bolstered by malicious users in a Sybil setting. In order to validate these propositions, we used the 500-node synthetic graph and identified a set of nodes, S_{sybil}, as *Sybil* by randomly selecting a seed and performing a breadth-first search until a number of attack edges are reached. This method of testing Sybil behavior is based on the technique in [42]. Throughout our experiments, this set of Sybil nodes are set to be highly interested in a topic T_y and have little interest in other topics whereas the interests of the other users are uniform among all the topics.



(a) Effect of Sybil p, q values. (b) Effect of # of Sybil nodes.

Figure 7: Sybil Attack Experiments.

We evaluated the relative importance of topic T_y which is the topic of interest of Sybil nodes, as a traditional, coordinated or uncoordinated trend with varying sizes of $|S_{sybil}|$. We answer two questions: 1) For a fixed size of Sybil nodes, how do the p and q values of Sybil nodes for T_y effect the relative trendiness of T_y as a traditional, coordinated and uncoordinated trend? 2) How does the size of Sybil nodes affect the same metric? In order to answer the former question, a set of experiments with increasing p and q values of Sybil nodes for T_y were performed where the Sybil attack size was set to 10 nodes. The results are presented in Figure 7(a) where the X-axis denotes the setting of the p and q values of Sybil nodes for topic T_y and Y-axis denotes the importance of T_y as a *traditional, coordinated* and *uncoordinated* trend. "Importance" refers to the number of topics T_y outranks (including T_y itself). So when T_y is the highest ranking topic, its importance is 50 as there are 50 possible topics in the data set. As it can be seen from Figure 7(a) with changing p and q values, coordinated score of T_y is consistently higher than traditional score and traditional score is higher than that of the uncoordinated score. It is also worthwhile to point out that, even with small values of p and q, we can see a breakpoint upon which T_y becomes significantly more trendy under coordinated trendiness, whereas this breakpoint is much later for the other two definitions.

A similar effect is observed in Figure 7(b) where the effect of the number of Sybil nodes in the trendiness of T_y is given. The X-axis refers to the number of Sybil nodes while the Y-axis demonstrates the same notion as the Y-axis in Figure 7(a). This set of experiments, for a fixed setting of p and q values (For Sybil nodes:$p_{i,y} = q_{i,j,y} = 0.9$ and $p_{i,k} = q_{i,j,k} = 0.01$ for $n_i \in S_{sybil}$, $T_k \in T - T_y$ and $n_j \in N$. As for non-sybil nodes: $p_{i,k} = q_{i,j,k} = 0.0.03$ for $n_i \notin S_{sybil}$, $T_k \in T$ and $n_i \in N$), tests the importance of the number of Sybil nodes and shows coordinated trendiness of T_y is consistently higher than its traditional and uncoordinated trendiness. Also, the jump in *coordinated* importance of T_y, which is useful for detecting the suspicious activity, can be observed with a small set of Sybil nodes whereas this jump is seen much later with the other definitions.

C.2 Further Results on Significance of Structural Trends

Here we provide some additional analysis and figures relating to the significance of *structural* trends in social networks. We first provide Table 6 relating to the model-based validation as given in Section 4.1. This table demonstrates the effect of increasing p val-

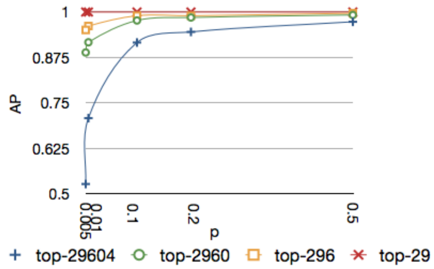


Figure 8: Average Precision of sampling technique for uncoordinated trends.

Table 6: Model Similarity Statistics.

p	$\rho_{trad-coor}$	$\rho_{trad-uncoor}$	AP_{coor}	AP_{uncoor}
0.1	0.763	0.988	0.144	0.571
0.3	0.518	0.993	0.079	0.672
0.5	0.401	0.996	0.062	0.737

Table 7: Coordinated trends in Twitter that are Traditionally Insignificant.

hashtag	R_{coor}	R_{trad}	hashtag category/explanation
#bb11	29	60	big brother 11 - tv series
#f1	35	70	sports
#green	33	74	political/social
#honduras	49	93	political/social
#ocra	44	91	Organized Conservative Resistance Alliance (political/social)
#digg	11	65	used mostly by digg.com users 70% retweet
#redsox	34	90	sports
#jesus	18	85	religious
#nieuws	27	97	news in Dutch
#hhrs	9	81	Hugh Hewitt Radio Show conservative talk show (political/social)

ues (probability that a node discusses a topic independent from its neighbors) in the similarity of *structural* and *traditional* trends. The results show that with increasing p values, the similarity between *traditional* and *uncoordinated* trends increases. The analysis summary is provided in Section 4.

We also provide two lists of Twitter hashtags in Tables 7 and 8 that list the details about 10 hashtags that have high *coordinated* and *uncoordinated* scores while having insignificant traditional rankings. These tables relate to the analysis-based validation of *structural* trend significance as discussed in Section 4.2. A large number of political/social/religious hashtags are listed in Table 7, whereas Table 8 mostly consists of idiom/generic concept hashtags. We further see that the list reported in Table 7 consists of narrow topics compared to the general topics in Table 8.

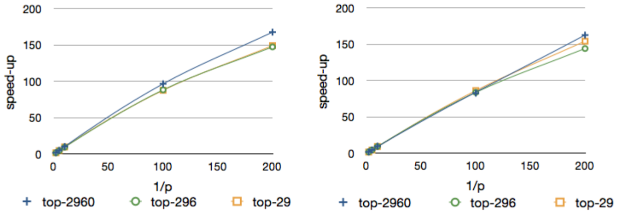
C.3 Further Results on Structural Trend Detection

In Section 5.3 we provided the figures that summarize the results of accuracy experiments for *coordinated* trends on the Twitter data set. As we noted before, the behavior of *uncoordinated* trends is similar to that of *coordinated* trends. We also noted that *uncoordinated* trends are more robust to sampling while providing a possible reasoning for this behavior. In this section, we provide the summary of the experiments for *uncoordinated* trends accuracy in Figure 8 for completeness. Similar to Figure 5, the X-axis denotes the rate of sampling whereas the Y-axis denotes the *average precision* for the given sampling ratio. Sampling ratio values used were $p = 0.5, 0, 2, 0., 0.01$ and 0.005. For top-29 topics, for all sampling ratios, we observe a perfect *average precision* of 1 while this value degrades rapidly for top-29604 *uncoordinated* topics.

Table 8: Uncoordinated trends in Twitter that are Traditionally Insignificant.

hashtag	R_{uncoor}	R_{trad}	hashtag category/explanation
#politics	6	83	political hashtag used by both conservatives and liberals
#marketing	15	92	generic concepts
#dontyouhate	19	96	idiom
#wheniwaslittle	7	84	idiom
#lol	17	94	generic concepts
#random	9	86	generic concepts
#5	10	88	hashtag related to being #5 an important aspect in Twitter
#twitterafterdark	4	82	idiom
#love	0	80	generic concept
#google	14	95	technology

As discussed in Section 5.3, sampling provides a linear speed-up. For completeness here we provide the figures that summarize the timing of experiments on the Twitter data set. Figure 9(a) provides the results for *coordinated* trends and Figure 9(b) provides the results for *uncoordinated* trends. For both figures, the X-axis denotes the inverse of sampling ratio $(1/p_s)$ and Y-axis is the speed-up, i.e. the ratio of the time it takes for the exact solution to the time it takes for sampling method to process the entire data set.



(a) Speed-up of sampling technique for coordinated trends.

(b) Speed-up of sampling technique for uncoordinated trends.

Figure 9: Speed-up of sampling.