Characterizing Social Influence in Google Buzz

Dallin Akagi Rishi Chandy Anthony Chong
Jonathan Krause Manuel Lagang

ABSTRACT

Google Buzz is a novel online service that presents new opportunities for social network analysis. By initializing the Buzz network with existing Gmail contacts, Google provides a unique dataset that may reflect a different aspect of online communication from those found in existing networks such as Facebook and Twitter. In this paper we design heuristic metrics for ranking and recommending influential members of the Buzz social network. We leverage these metrics to develop an application allowing individual Buzz users to identify influential users near their existing "friend" subgraph.

Categories and Subject Descriptors

J.4 [**Computer Applications**]: Social and Behavioral Sicences—*Sociology, Economics*; H.2.8 [**Database Applications**]: Data mining; H.3.3 [**Information Search and Retrieval**]: Text Mining

General Terms

Economics, Measurement, Human Factors, Experimentation

Keywords

social influence, social networks, google buzz

1. INTRODUCTION

Social networks, i.e. graphs of the relationships between a group of individuals, provide a fundamental tool in understanding how ideas propagate among people. Such graphs have been used to analyze various topics from how the Medici family gained power in Renaissance Florence [24] to the dynamics of friendships and romances in high school students [2]. Common in the sociologist's treatment of social networks are the metrics of node centrality. In particular, measuring in-degree, betweenness, and eigenvector centrality are common practice [4]. Determining which of these metrics to use on a particular dataset are generally based on heuristics.

Online social networks have evolved into a rich setting for social network analysis. Various authors have discussed the degree of influence and privacy in networks like Facebook, MySpace, and Twitter [21, 23, 9]. However, the network research applied to Google Buzz[1] remains limited.

In February 2010, Google deployed Buzz, its social networking and messaging tool, with user profiles linked to all existing Gmail accounts [12]. This provides a substantive framework for social network analysis, since Buzz may reflect existing relationships found in email communication. Furthermore, the multidimensional nature of the data available on Buzz provides an interesting dataset for analysis: Users "follow" one another, creating a follower-followee graph (a type of "friend graph"). Additionally, they may indicate that they "like" another user's post, and comment on posts they find interesting, presenting unique challenges in choosing and blending the best influence metrics for each component to arrive at an overall influence score for every user. Our research aims to extend existing methods, implement new metrics for social influence, and evaluate performance.

1.1 Previous Literature

Prior research focuses on the influence maximization problem in social networks[27, 26]. In order to characterize the dynamics of viral marketing, Kempe, Kleinberg, and Tardos[13] attempt to determine social influence by asking: If we can try to convince a subset of individuals to adopt a new product or innovation, and the goal is to trigger a large cascade of further adoptions, which set of individuals should we target? They model network influence through diffusion models (namely the Linear Threshold and Independent Cascade Models) on social networks. By applying a Domingos-Richardson [6] style of optimization, Kempe, et al. were able to create an algorithm that significantly outperforms traditional node-selection heuristics based on distance and degree centrality in identifying influential agents in the physics coauthorship graph on arXiv. However, the gradient ascent (greedy hill climbing) method they utilize requires the use of the n-dimensional gradient, which involves intensive computation.

A different treatment of influence maximization problems considers the similarity to disease outbreak problems. Kimura, Saito, and Nakano [15] introduced a more efficient technique to evaluate these models based on graph theoretic optimizations. These models have been experimentally evaluated on

[1] http://www.google.com/buzz

a large sample of blog "trackback" data and on a maximal connected component of people mentioned on Wikipedia.

Building on the idea of peer influence, Tang, et al. [27] analyzed the topical influence of individuals in social networks. They propose Topical Affinity Propagation (TAP) to model the importance of topic-level influence propagation. In particular, they seek to determine the representative nodes on a topic and how to determine social influence of neighboring nodes of a particular node. TAP is based off of the theory of a factor graph [8] in which observational data is coupled with local attributes and connections. By leveraging affinity propagation in this setting, Tang, et al. are able to create a model for influence identification through two different methods: Topical Factor Graph (TFG) and TAP Learning (and distributed TAP learning). Experiment results confirm the success of TAP in identifying topic-based influence in real-life large data sets. Additionally, the distributed learning model proves to be scalable with reasonable performance.

On a related topic, Bharathi, et al.[26] discuss the effect of social networks on the diffusion of ideas and innovation. Similar to Kempe, et al., Bharathi, et al. provide an approximation algorithm to computing the best response to an opponent's strategy in the "game of innovation". Specifically, we again consider the idea of activated nodes. In the influence maximization game, players wish to maximize their individual influence given a randomized propagation scheme. It can be shown that mixed Nash Equilibria exist for this game (but no pure Nash Equilibrium). From here, Bharathi, et al. show that best-response strategies exist for this game that are both monotone and submodular. This, coupled with discussion of "first mover" strategies provides a framework for the behavioral basis of influence maximization in social networks.

An interesting phenomenon of influence is an "information cascade", in which individuals adopt a new idea based on the influence of others. Leskovec, Singh, and Kleinberg [20] provide an analysis of this concept on social networks by looking at the cascading effect of recommendations. Extending the previous work of sociologists who looked at the "diffusion of innovation" [25] to an online setting, they seek to characterize the nature and scope of these cascades. By conducting their analysis on a peer-to-peer recommendation network consisting of 4,000,000 users and 16,000,000 recommendations on 500,000 products, they found that the distribution of cascade sizes is heavy-tailed. Cascade patterns were found to be generally shallow and tree-like subgraphs, with patterns not directly related to size or intensity, which suggests that cascading behaviour is dominated by underlying network properties.

1.2 Google Buzz

At the most basic level, Google Buzz allows users to post messages to their activity streams. They can also interact with others' posts by commenting on them or "liking" them, which adds their name to a list of "likers." Unlike Twitter, there is no limit to the length or type of content that a post may contain.

1.2.1 Follower-Followee Model

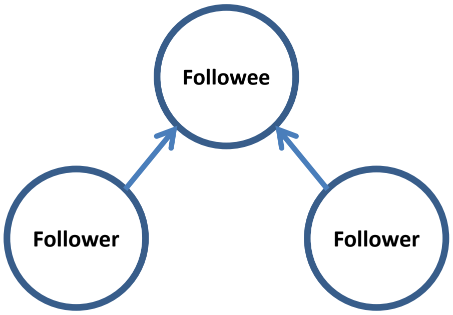

Figure 1: Follower-Followee Graph Model

The dynamics of the Buzz social graph are very similar to those present in Twitter. In the follower-followee model (Figure 1), if user A is following user B, then there is a directed edge from user node A to user node B. By counting the number of times user A "likes" posts by user B, along with other metadata counts, we can compute weights for the edges in this graph. Social influence travels along reverse edge direction, with the exception of "likes" and comments.

2. APPROACH

In this section we present the details of our approach to data collection and social influence analysis.

2.1 Data Collection

The graph structure of the Google Buzz network is so vast that it is infeasible to analyze in its entirety. Thus, a subgraph from the network was sampled in order to get a representative view on the general structure. The sampling method chosen is similar to a breadth-first search, but incorporates randomness by choosing the order of expanding nodes regardless of distance from the seed node. The pseudocode for the sampling method is shown in Algorithm 1.

Algorithm 1 POOL-SAMPLE

```
POOL = V0 {V0 is the seed node}
while POOL $\neq \emptyset$ do
    V = Uniform random selection from V0
    POOL = POOL \ {V}
    Sample data for V
    POOL = POOL $\cup$ neighbors of V
end while
```

2.1.1 Buzz Dataset

According to Google, there are "millions" of Buzz users, each with multiple follower-followee relationships with other users. In order to test our methods and develop a prototype, we created a sample dataset by crawling 41,858 user profiles involving 204,289 relationships with 3,394,137 Buzz posts. We also crawled all comments and "likes" among the users in the sample dataset. Figure 2 shows that most Buzz posts actually originate from Twitter. Still, the extra functionality that Buzz provides over Twitter, including direct commenting on posts and "liking," could be valuable.

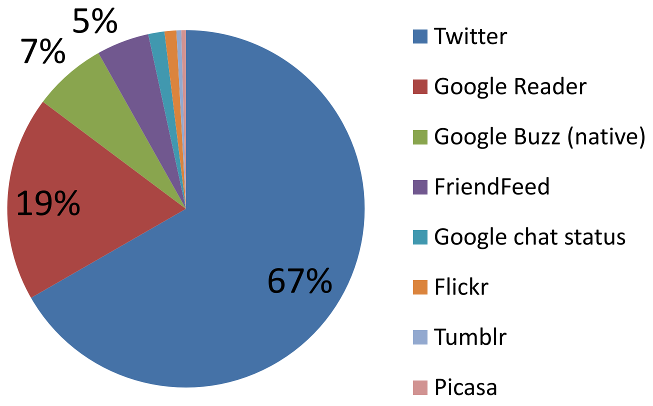

Figure 2: Pie chart of the source of Google Buzz posts (from sample dataset)

2.2 Sampling Bias

Because of the large size of the online social networks we are studying, practical considerations prevent us from crawling the entire graph. Instead, we utilize the common approach of collecting and analyzing a smaller, representative sample of the network. Collecting a relatively small sample of a vast network necessitates analyses of biases introduced by the sampling method.

Empirical observations[22][18][3] suggest that Breadth First Search systematically favors higher degree nodes, while random walks perform well in choosing representative subgraphs[19]. POOL-SAMPLE combines the best of both methods, and, to our knowledge, has not been analyzed to much extent. We intend to characterize the bias of our sampling method and offer suitable corrections. Fairly little analysis has been done on the sampling bias of searches on online social networks. The most notable results are from Kurant et al[17], which we consider here.

For a given degree distribution p_k we can generate a random graph $RG(p_k)$ from which to sample. In this setting, Breadth First Search is equivalent to other graph traversal techniques such as Depth First Search, Snowball Sampling, and Forest Fire. Furthermore, the bias from Breadth First Search is identical to the bias from Random Walk. This bias is monotonically decreasing with an increasing fraction of sampled nodes f. However, even given a biased sample, we can give an unbiased estimator of the original degree distribution:

$$\widehat{p_k} = \frac{\widehat{q_k}}{1-(1-tf)^l} \cdot \left(\sum_l \frac{\widehat{q_l}}{1-(1-tf)^l} \right)^{-1}$$

Here, q_k is the distribution observed (biased towards high-degree nodes), at time t.

2.3 Influence Metrics

Influence is difficult to describe, much less quantify into numerical values for each node. We chose several metrics for analysis, with the criteria that they capture some intuitive notion of influence. Each of these metrics map a node i to a value $I_i \in \mathbb{R}$.

- **In-Degree** (ID) The size of the set of nodes that have an edge leading to i. It is natural to believe that a person with a large amount followers is influential.
- **In-Web**$< k >$ (IW) A generalization of Indegree. This counts the number of nodes that have a directed path to i of length at most k.
- **H-Index** (HI) The H-Index was proposed by Hirsch [11] as a means to quantify an individual's scientific output based on the structure of the citation graph. The integer score is a count of the number n of papers written by an individual which have each been cited at least n times. It requires that an individual have a large number of highly cited papers in order to improve his or her index rating, and lessens the impact of a single highly-cited paper. However, the H-Index has several drawbacks which we do not discuss here because they are more relevant for the validity of measuring a scientist's impact than its validity as a graph metric. Nonetheless, it has seen wide implementation as a metric of an individual's scientific output.

  We have adapted the H-Index for use in social networks with directed graphs. An individual's followers are seen as a parallel to publications, and the respective followers of those followers are seen as a parallel to citations. Hence, if an individual has 50 followers who each have at least 50 followers themselves, he or she would have an H-Index of 50.

  This seems to be a valid metric of the capability to influence others because it corresponds to high connectedness. It also conveys more information than In-Degree because it contains the notion of being able to influence highly influential people. As an added benefit, it can also be computed efficiently using only local data.
- **Random Walk** (RW) This metric measures the time-average probability of being on node i during a random walk. Random walk models have been used in PageRank to measure authority of Internet pages.

  Our implementation of Random Walk is as follows: For some specified number of iterations, pick a random node to start from. Then, proceed with a random walk by random choosing amongst the out-edges of the current node and continuing the random walk at that node. In each iteration, with a specified probability, restart the random walk.

  Alternatively, one can get the matrix for a Markov chain determined by this random walk and solve for the eigenvalues of the matrix to determine metric values, but when there are tens of thousands of nodes, this computation is too slow. This formulation is equivalent to the explicit random walking as the Markov chain determined by the graph structure imposed is ergodic.

- **Independent Cascade Diffusion** (IC) Diffusion models have been used to analyze the ability of a node to infect the network, particularly for targeted viral marketing. The independent cascade model probabilistically activates edges to propagate infection, and Monte Carlo samples are used to measure the expected size of the infected set.

2.3.1 Personalization With Local Influence

Thus far, we have concentrated on the task of measuring the influence of users in a global context. However, for the task of recommendation targeted for specific users, the concept of global influence becomes less important. Users may be more concerned with influential nodes relative to themselves, thus a measure of local influence must be devised.

A natural way of localizing metrics is to restrict the measurement process to a local subgraph. This restriction can be done in several ways: measure the global influence of all nodes and only recommend the highest nodes in the local subgraph, or use only the local subgraph to measure the influence of local nodes. However, restricting decisions to an arbitrary local subgraph in this manner is sub-optimal as much of the information in the graph is unused. Also, many of the metrics used are local in nature (In-Degree, InWeb, H-Index), thus restricting recommendations to a large local subgraph will still be similar to picking globally influential nodes regardless of target user.

The method of recommendation we have used measures a non-local metric (Random Walk) on a slightly modified graph to target a particular user. The modification to the graph involves adding extra edges that are implicitly present in a random walk to ensure that the underlying Markov chain is ergodic. Many applications have these extra edges connecting every pair of nodes with equal weights such that the sum of the weights from any node to any of these edges is α. For personalization, these edges are allowed only to go into the local subgraph centered around the target node. This no longer ensures ergodicity of the Markov chain of the whole graph as the graph may be unconnected. However, the chain consisting only of nodes reachable from the local subgraph is ergodic, so random walks restricted to this chain will give a convergent solution.

This method enables the use of information present in the whole graph while localizing the measure of influence for a target user. This method also captures an intuitive meaning of localized influence: if information tends to pass from followee to follower, what are the nodes that can pass the most information to nodes around the target? In practice, this method seems to be acceptable: many recommendations are not in the global top leaderboards, often recommended before those that are. However, without a way to validate the results, we cannot completely justify the validity of this method for recommendation purposes.

3. RANK AGGREGATION

Our aim was to create a single metric which could be applied to a social network to give consistent friend recommendations containing the most influential users in the network. In order to aggregate the multidimensional features that contribute to the notion of influence, we have selected several metrics which capture different aspects of social influence. We utilize a method of aggregating various ranking metrics in order to succinctly represent the collective body of information contained in these metrics. Additionally, we sought to produce a ranking system resistant to attack from users seeking to artificially inflate their rankings, and also easily computable, thus allowing the ranking to be queried on-demand in a dynamic social network.

Dwork et al [7] propose a method for aggregating web search engine results in a robust manner which protects users from various shortcoming and biases in the various search results. We use their method for rank aggregation which benefits from having the criteria that we sought to establish. We also evaluate the shortcomings of the method and discuss some possible enhancements.

There are two broad steps we implement to arrive at a rank aggregation which has the benefits described above. The first step is rank aggregation via Borda's method. The second step is rank refinement by adjacency swaps on the aggregate.

3.1 Borda's Method

The Borda count is an election method in which voters rank candidates in order of preference. In terms of rank aggregation each ranking system used is seen as a voter and each member of the set is a candidate. Scores are assigned to each rank and each member's final score is the sum of their scores from the various ranking metrics.

Formally, given full lists $\tau_1, \tau_2, ..., \tau_k$ for each candidate $c \in S$ and list τ_i, Borda's method assigns a score

$$B_i(c) = |\{x \in \tau_i : \text{x ranked worse than } c \text{ in } \tau_i\}|$$

and then the total Borda score for the candidate is

$$B(c) = \sum_{i=1}^{k} B_i(c)$$

The candidates are then sorted in decreasing order of total Borda score.

3.2 Rank Refinement

One widely accepted metric for concordance amongst various rankings is the Kendall distance. Kemeny optimal aggregations, i.e. those that optimize Kendall distance, have been shown to be unique aggregates which are neutral, consistent, and which satisfy the Condorcet criterion. Computing the Kemeny optimal aggregation has been shown to be NP-Hard [7].

In order to arrive at a tractable result, we follow the method for local Kemenization proposed by Dwork et al [7]. Given the ranked lists $\tau_1, \tau_2, ..., \tau_k$ and the aggregate σ, we attempt to swap adjacent entries in σ which will lower the Kendall distance on the whole collection of rankings:

$$K(\sigma, \tau_1, \tau_2, ..., \tau_k) < K(\sigma', \tau_1, \tau_2, ..., \tau_k)$$

3.3 Benefits and Consequences of Aggregation

The method described above produces a ranking which satisfies the extended Condorcet criterion, i.e. if a majority of rankings position x above y, then x is ranked above y in the final ranking. In such a procedure it is more difficult for one member to try to artificially inflate his or her ranking via spam or automated action. Thus the ranking is useful for users because it establishes a level of trust.

Additionally, the above method can be computed efficiently once the ranking lists have been computed. This allows for the ranking to be utilized on social networks whose structure and activity changes frequently while still conveying useful information.

The rank refinement acquired by arriving at a locally Kemenized list is limited by the original aggregation. It is in a sense maximally consistent with the original aggregate, and so cannot arrive at a final ranking which will convey useful information if the original ranking was poorly determined.

Borda's method gives equal amount of importance to every ranking system. This may not be desirable in a social network, and could allow some members to be misrepresented in the final standings. However, connectivity and activity in a social network are both major factors in determining influence and hence our aggregate captures that notion well. It remains to be seen whether some linear combination of the points assigned in Borda's method (a weighting) would give results which are more consistent with intuitive expectations.

4. METRIC COMPARISONS

4.1 Kendall's Tau

Comparing different influence metrics is equivalent to comparing the rankings that they impose on our social network. To that end, we compared metrics using the Kendall's tau coefficient [14]. The Kendall's tau coefficient is defined as

$$\tau = \frac{\sum_{(i,j)}[(i,j) \text{ in same order}] - \sum_{(i,j)}[(i,j) \text{ in different order}]}{n(n-1)}$$

where n is the total number of nodes and the sums are being taken over all pairs of nodes. Note that $\tau \in [-1,1]$, with $\tau = 1$ corresponding to complete ranking agreement between the metrics, $\tau = -1$ corresponding to complete disagreement, and $\tau \approx 0$ corresponding to no relation whatsoever.

We performed metric comparisons using both the Google Buzz and StackOverflow datasets in order ensure that our comparisons are valid. The results are in Figures 1 and 2. For these results, Random Walk was done with 1 billion walks and probability 0.2 of starting a new walk at any given step, and Independent Cascade was done with 100 trials per node, with an activation probability of 0.1. Additionally, for the StackOverflow dataset we included an additional metric, User Reputation (abbreviated REP), which simply uses a user's public reputation score on StackOverflow, which should capture the notion of influence in that network.

Looking at the tables, all of the numbers in Table 1 and Table 2 are positive, which indicates that the metrics are roughly measuring similar things, a good sanity check. Looking at the Google Buzz data in particular, some items stand out. Hirsch Index is very similar to In-Degree, which is expected due to the definition of the Hirsch Index. Independent Cascade is similar to the In-Web metrics, which is valid as In-Web is basically Independent Cascade with probability of activation 1 and limited view a certain distance away from the node under consideration. Both of the In-Web metrics are also similar to each other, which is completely expected. However, nothing is very similar to Random Walk. Note that this does not necessarily mean that Random Walk is a bad metric; it just means that it is different from the other metrics presented.

Table 1: Kendall's Tau Coefficients for Buzz Dataset

	HI	IC	ID	IW(2)	IW(3)	RW
HI	1.000	.2665	.8122	.2689	.2125	.0868
IC	.2665	1.000	.3645	.7823	.8140	.1382
ID	.8122	.3645	1.000	.3645	.3090	.2411
IW(2)	.2689	.7823	.3645	1.000	.8349	.1056
IW(3)	.2125	.8140	.3090	.8349	1.000	.1021
RW	.0868	.1382	.2411	.1056	.1021	1.000

Table 3: Difference in Kendall's Tau

	HI	IC	ID	IW(2)	IW(3)	RW
HI	0.000	.1256	-.1370	.1265	.1647	.1337
IC	.1256	0.	.2319	.0896	.0602	.3576
ID	-.1370	.2319	0.	.2331	.2706	.2264
IW(2)	.1265	.0896	.2331	0.	.0950	.3812
IW(3)	.1647	.0602	.2706	.0950	0.	.3822
RW	.1337	.3576	.2264	.3812	.3802	0.

Now focusing on the StackOverflow dataset, we notice similar patterns. More explicitly, we can take the difference in Kendall's tau coefficients, as in Table 3. From this, we can see that the Kendall's tau coefficients for the StackOverflow dataset are on average approximately 0.2010 different in absolute value and 0.1828 higher on average. Very significantly, the StackOverflow coefficients are almost uniformly higher than the Google Buzz coefficients, with the only exception being H-Index against In-Degree. Also, more of the metrics are similar to Random Walk as compared to the Buzz dataset, although the correlation with Random Walk is still not as high as the other correlations.

However, the User Reputation metric is very different from all of the other metrics. If User Reputation were the definitive and ultimate social influence metric on StackOverflow, then this would indicate that none of our presented metrics are a good measure of influence, assuming that the graph structure we imposed on the StackOverflow dataset was valid.

4.2 CCDFs

Now we present some complementary cumulative distribution functions (CCDFs) on a log-log scale, noting that many real-life distributions are heavy-tailed and thus have linear CCDFS when plotted on a log-log scale. There is no particular reason to believe that some of these metrics are linear, though, but it is worth investigating. For reasons of space, though, not all CCDFs have been included.

In Figure 3, we can see that In-Degree on the Google Buzz set is somewhat linear, which is more or less expected. In Figure 4, the same general trend can be observed for Stack-

Table 2: Kendall's Tau Coefficients for StackOverflow Dataset

	HI	IC	ID	IW(2)	IW(3)	RW	REP
HI	1.000	.3921	.6752	.3954	.3772	.2205	.0863
IC	.3921	1.000	.5964	.8719	.8742	.4958	.2749
ID	.6752	.5964	1.000	.5976	.5796	.4675	.2118
IW(2)	.3954	.8719	.5976	1.000	.9299	.4868	.2616
IW(3)	.3772	.8742	.5796	.9299	1.000	.4843	.2597
RW	.2205	.4958	.4675	.4868	.4823	1.000	.2597
REP	.0863	.2749	.2118	.2616	.2597	.2597	1.000

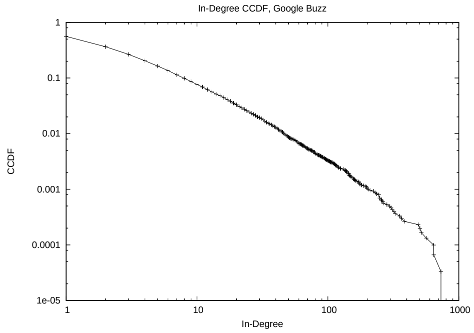

Figure 3: CCDF of In-Degree on Google Buzz dataset.

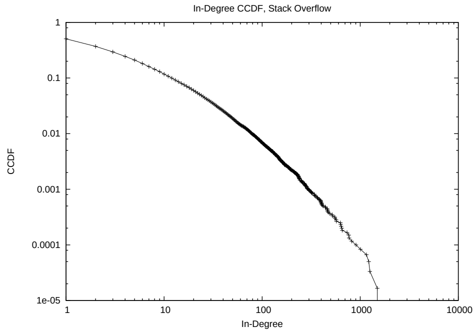

Figure 4: CCDF of In-Degree on StackOverflow dataset.

Overflow, indicating that our crawling techniques are not terribly biased. To illustrate a metric that is not linear, observe the CCDF for In-Web(3) on the Google Buzz set (Figure 5). StackOverflow Reputation (Figure 6) is also non-linear on a log-log scale.

In addition to the above observations, although the plots are not presented, the CCDF curves for all of the metrics are fairly similar between Google Buzz and StackOverflow. This furthermore indicates that these datasets are not very different and validates our imposition of graph structure on the Stack Overflow dataset.

However, it is also important to note that merely having similar shapes does not make metrics similar. For example the CCDF of In-Web(2) on StackOverflow looks quite similar to the CCDF of StackOverflow Reputation, yet the Kendall's tau coefficient for these two metrics is only .2616.

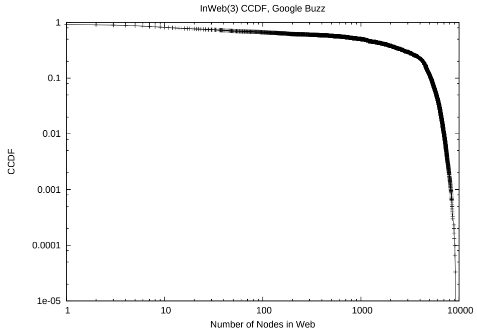

Figure 5: CCDF of In-Web(3) on Google Buzz dataset.

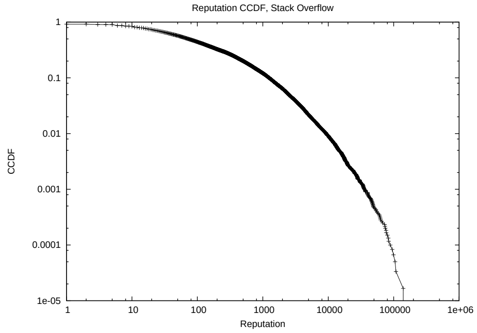

Figure 6: CCDF of StackOverflow Reputation.

5. EVALUATION

Previous literature in social influence analysis focuses on analyzing new or existing metrics, while ignoring the problem of evaluating their effectiveness. This is due to the difficulty in finding an appropriate test dataset labelled with pre-determined social influence scores. We decided to evaluate our methods using a relatively new dataset [16] derived from the StackOverflow online question and answer website.

5.1 StackOverflow Dataset

StackOverflow provides a dataset containing 227,691 users, 2,488,534 posts, and 6,444,449 individual votes. We imported this into a MySQL database using a custom PHP script. Users of StackOverflow can vote on or "favorite" questions posted by other users. To preserve user privacy, votes are omitted from the public dataset. Based on various criteria, each user has a public "reputation score" which we use as labels for users' relative influence. In order to derive a graph analogous to the follower-followee model, we created

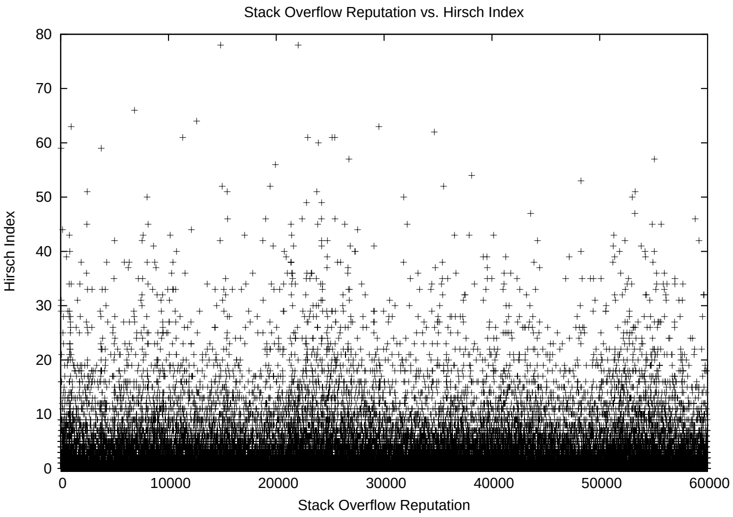

Figure 7: Scatterplot of StackOverflow Reputation Score vs H-index

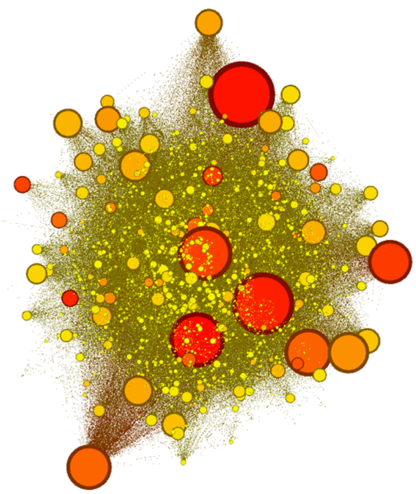

Figure 8: Buzz graph visualization, based on In-degree (node size) and Pagerank (node color; red = high, yellow = low)

a directed edge from user A to user B if user A marks a question posted by user B as one of their favorites. This graph contains 377,780 directed relationships.

5.2 Evaluation Results

Analysis of the relationship between H-index scores and "reputation score" labels shows that there is no significant correlation between these two metrics (Figure 7). There are many reasons for this, the most intuitive being that StackOverflow is not a social service in the same way that Buzz or Twitter are inherently social services. StackOverflow is primarily meant for answering questions posed by users, and thus any social aspects are merely secondary effects. The notion of "following" in actual social networks is stronger than in the graph we derived from the StackOverflow dataset.

6. FUTURE WORK

Chen, et al. [5] point out the infeasibility of running the greedy algorithm proposed by Kempe, et al. [13] on very large datasets and puts forward degree discount heuristics which provide comparable results with computation time 6 orders of magnitude faster. We aim to investigate whether the same speed-ups are necessary in the restriction of the problem to an ego-network subgraph. In addition, we would like to evaluate our method of friend recommendations to see how they perform with respect to degree centrality metrics.

There are many technical issues associated with large-scale data analysis, including efficient data storage and quick retrieval. In the future, we plan to investigate the possibility of using graph databases, such as neo4j, to allow easier access to the follower-friend graph. This would allow us to expand the size of the dataset without sacrificing performance.

As mentioned earlier, evaluation is an important aspect of social influence analysis that has been neglected in previous work. Although StackOverflow proved unsuitable as a test dataset, in the future we aim to evaluate our methods using an appropriate labelled dataset.

On May 27, 2010, Google released a new "reshare" feature for Buzz which allows users to copy others' posts into their own Buzz activity stream, similar to the "retweet" feature of Twitter. These copied posts retain a link to the orginal poster, and it is possible to have a chain of "reshares." So far, our analysis has been limited to working with just the follower-followee graph and related metadata. Now, we can also take into account these "reshare" chains, similar to the established "information cascades" models. Bakshy, et al. [1] had access to a dataset in which adoption of new content was readily perceivable and was thus able to observe actual cascades of influence rather than the simple potential for influence. We would like to correlate our recommendations based on network structure with empirical results such as these to determine whether network structure alone is effective in locating influential nodes. This is even more relevant because they come from what Guo, et al. [10] call "networking oriented" online social networks, and are categorized as those where content sharing is mainly among friends, and where the networks are driven by the underlying social relationships.

7. PROJECT CHALLENGES

7.1 Initial Crawler

One of the initial challenges was properly crawling a subgraph. The first method attempted was the basic breadth first search, keeping nodes in a FIFO queue in the order discovered and expanding all the neighbors of the head of the queue. This leads to a quick explosion of the number of nodes and edges in the sampled graph; in just 1500 expanded nodes, the number of total nodes grew to more than 200,000 and the number of edges numbered more than one million.

This explosion has many implications for the quality of the sampled subgraph. The rapid growth of nodes upon expanding implies that when limiting the overall size for practicality purposes, the distance of the nodes expanded to the initial seed is very small, so intuitively an unfair bias is present toward the initial seed node. Also, many of the nodes in the graph are "leaves", nodes that are were not expanded and thus likely have degree 1, so the majority of the graph provides little information. These issues were hopefully addressed by the randomized pool sampling algorithm presented previously in the paper.

7.2 Code Performance and Data Management

Significant effort went into choosing the most efficient data storage configuration. Although we eventually decided on a traditional MySQL relational database, we also explored the possibility of using specialized graph databases (such as Twitter's FlockDB). Unfortunately, they proved to be too slow on our hardware.

In addition, crawling performance was also an issue. Crawling too quickly would result in Google blocking our machine, but crawling too slowly would mean the crawl would take an inordinate amount of time. We were also limited in the rate at which several threads could access the shared queue of URLs without blocking.

8. ACKNOWLEDGMENTS

We thank Professor Steven Low, Professor Adam Wierman, and Minghong Lin for their valuable guidance and mentorship for this project.

9. REFERENCES

[1] E. Bakshy, B. Karrer, and L. A. Adamic. Social influence and the diffusion of user-created content. In *EC '09: Proceedings of the tenth ACM conference on Electronic commerce*, pages 325–334, New York, NY, USA, 2009. ACM.

[2] P. S. Bearman, J. Moody, and K. Stovel. Chains of affection: The structure of adolescent romantic and sexual networks. *The American Journal of Sociology*, 110(1):44–91, 2004.

[3] L. Becchetti, C. Castillo, D. Donato, A. Fazzone, and I. Rome. A comparison of sampling techniques for web graph characterization. In *Proceedings of the Workshop on Link Analysis (LinkKDD'06), Philadelphia, PA*, 2006.

[4] R. S. Burt and M. J. Minor. *Applied Network Analysis: A Methodological Introduction.* Sage Publications, Beverly Hills, 1983.

[5] W. Chen, Y. Wang, and S. Yang. Efficient influence maximization in social networks. In *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 199–208, New York, NY, USA, 2009. ACM.

[6] P. Domingos and M. Richardson. Mining the network value of customers. In *KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 57–66, New York, NY, USA, 2001. ACM.

[7] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation methods for the web. In *WWW '01: Proceedings of the 10th international conference on World Wide Web*, pages 613–622, New York, NY, USA, 2001. ACM.

[8] B. J. F. Frank R. Kschischang and H.-A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47(2):498–519, February 2001.

[9] R. Gross and A. Acquisti. Information revelation and privacy in online social networks. In *WPES '05: Proceedings of the 2005 ACM workshop on Privacy in the electronic society*, pages 71–80, New York, NY, USA, 2005. ACM.

[10] L. Guo, E. Tan, S. Chen, X. Zhang, and Y. E. Zhao. Analyzing patterns of user content generation in online social networks. In *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 369–378, New York, NY, USA, 2009. ACM.

[11] J. E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46):16569–72, November 2005.

[12] M. A. Kaafar and P. Manils. Why spammers should thank google? In *SNS '10*, New York, NY, USA, 2010. ACM.

[13] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146, New York, NY, USA, 2003. ACM.

[14] M. Kendall and J. Gibbons. *Rank Correlation Methods.* Charles Griffin, 5th edition, 1990.

[15] M. Kimura, K. Saito, and R. Nakano. Extracting influential nodes for information diffusion on a social network. In *AAAI'07: Proceedings of the 22nd national conference on Artificial intelligence*, pages 1371–1376. AAAI Press, 2007.

[16] R. Kumar, Y. Lifshits, and A. Tomkins. Evolution of two-sided markets. In *WSDM '10: Proceedings of the third ACM international conference on Web search and data mining*, pages 311–320, New York, NY, USA, 2010. ACM.

[17] M. Kurant, A. Markopoulou, and P. Thiran. On the bias of bfs. Apr 2010.

[18] S. H. Lee, P.-J. Kim, and H. Jeong. Statistical properties of sampled networks. *Phys. Rev. E*, 73(1):016102, Jan 2006.

[19] J. Leskovec and C. Faloutsos. Sampling from large graphs. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 631–636, New York, NY, USA, 2006. ACM Press.

[20] J. Leskovec, A. Singh, and J. Kleinberg. Patterns of influence in a recommendation network. In *In Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, pages 380–389. Springer-Verlag, 2005.

[21] K. Lewis, J. Kaufman, M. Gonzalez, A. Wimmer, and N. Christakis. Tastes, ties, and time: A new social network dataset using facebook.com. *Social Networks*, 30(4):330 – 342, 2008.

[22] M. Najork and J. L. Wiener. Breadth-first crawling

yields high-quality pages. In *WWW '01: Proceedings of the 10th international conference on World Wide Web*, pages 114–118, New York, NY, USA, 2001. ACM.

[23] A. Nazir, S. Raza, and C.-N. Chuah. Unveiling facebook: a measurement study of social network based applications. In *IMC '08: Proceedings of the 8th ACM SIGCOMM conference on Internet measurement*, pages 43–56, New York, NY, USA, 2008. ACM.

[24] J. F. Padgett and C. K. Ansell. Robust action and the rise of the medici. *The American Journal of Sociology*, 98(6):1259–1319, May 1993.

[25] E. Rogers. *Diffusion of Innovations.* Free Press, 5 edition, 2003.

[26] D. K. Shishir Bharathi and M. Salek. Competitive influence maximization in social networks. In *Lecture Notes in Computer Science*, pages 306–311, Berlin, Germany, 2007. Springer Berlin / Heidelberg.

[27] J. Tang, J. Sun, C. Wang, and Z. Yang. Social influence analysis in large-scale networks. In *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 807–816, New York, NY, USA, 2009. ACM.