Nature of Information, People, and Relationships in Digital Social Networks

Rakesh Agrawal
Microsoft Research
Mountain View, CA 94043, U.S.A.
rakesha@microsoft.com

Abstract

This paper summarizes the results of our recent investigations into how information propagates, how people assimilate information, and how people form relationships to gain information in Internet-centric social settings. It includes key ideas related to the role of the nature of information items in information diffusion as well as the notion of receptivity on part of the receiver and how it affects information assimilation and opinion formation. It describes a system that incorporates availability, willingness, and knowledge in recommending friends to a person seeking advice from social network. It discusses whether having common interests makes it more likely for a pair of users to be friends and whether being friends influences the likelihood of having common interests, and quantifies the influence of various factors in an individual's continued relationship with a social group. Finally, it gives current research directions related to privacy and social analytics.

1 Introduction

The mission of Microsoft Research's Search Labs in Silicon Valley that I lead is to advance the state of art in Internet technologies and Internet-based applications. One of our focus areas is to understand how information propagates, how people assimilate information, and how people form relationships to gain information in Internet-centric social settings. This paper presents a condensed overview of some of our recent research on these topics. It includes key ideas related to the role of the nature of information items in information diffusion, presented by Agrawal, Potamias, and Terzi in [1]. It also discusses the notion of receptivity on part of the receiver and how it affects information assimilation from the same paper. Related to the same topic, it introduces the work of Bhawalkar, Gollapudi, and Munagala on opinion formation games from [2] and that of Das, Gollapudi, Panigrahy, and Salek on dynamics of opinion formation from [5]. It then reviews the system of Nandi, Paparizos, Shafer, and Agrawal that factors in availability, willingness, and knowledge to recommend friends for person to turn to for advice. Next, it recalls the work of Lauw, Shafer, Agrawal, and Ntoulas from [6] to shed light on whether having common interests makes it more likely for a pair of users to be friends, and whether being friends influences the likelihood of having common interests. Finally, it abstracts the work of Budak and Agrawal from [3] on factors that influence an individual's continued relationship with a social group. The work

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

in [3] is based on data from thirty Twitter chat groups; algorithmic mining of chat groups from Twitter stream is described by Cook, Kenthapadi, and Mishra in [4].

Mea Culpa: Given the space restriction, I have prioritized the presentation of Search Labs work over discussion of related research. For the latter, I refer the reader to the original papers.

2 Nature of Information

A key issue in social networks is understanding how people assimilate information in their daily lives. Recent research has focused on understanding the role that *node characteristics* (i.e., homophily) and *peer influence*, (i.e., link structure), play in explaining the appearance of information items on certain nodes of the social network. The underlying assumption is that it is the nature of the people, or the nature of the people's connections, which determines the form of information cascades.

While we recognize the importance of network structure and nodes' characteristics on information propagation, we postulate in [1] that the very nature of information items is an additional important parameter that affects the observed spread. We claim that certain information items are *endogenous* and they indeed propagate primarily through the connections between the nodes. On the other hand, some information items are *exogenous* – they will be acquired by many nodes independently of the underlying network. Given a social network and data related to the ordering of adoption of information items by nodes, our goal is to develop a framework for estimating endogeneity and exogeneity parameters.

E2 **Model:** Consider a social network $G = (V, E)$ of $|V| = n$ users, in which there is a link $(u \rightarrow u')$ between two nodes $u, u' \in V$, if node u *follows* node u'. Such a directed link suggests that there is potential of information propagation from u' to u. Assume a finite set of information items $\mathcal{I}$ with $|\mathcal{I}| = m$.

At every point in time t, every node $u \in V$ is associated with an m-dimensional vector A_u^t, whence $A_u^t(i) = 1$ if node u is *active* with respect to information item i at time t; otherwise $A_u^t(i) = 0$. If $A_u^{(t-1)}(i) = 0$ and $A_u^t(i) = 1$, then we say that an *activation* has occurred to node u with respect to item i at time t. The *observed* activation state at the end of the observation period is encoded in $\mathbf{A}$ such that $\mathbf{A}(u, i) = 1$ iff node u has, at some point, become active with respect to item i. Give the sequence of activations encoded in vectors A_u^t, one can construct the *active-neighborhood* matrix $\mathbf{\Gamma}$, such that $\mathbf{\Gamma}(u, i)$ denotes the number of neighbors of u that were active with respect to item i, the moment u became active with respect to i. If $\mathbf{A}(u, i) = 0$, then $\mathbf{\Gamma}(u, i)$ is the number of neighbors of u that were active at the end of the observation period.

Every item $i \in \mathcal{I}$ is characterized by a pair of parameters $\theta_i = (e_i, x_i)$, where $e_i \in [0, 1]$ is its *endogeneity* and $x_i \in [0, 1]$ is its *exogeneity*. Endogeneity characterizes the item's tendency to propagate through the network due to the peer effect. Exogeneity captures the item's tendency to be independently generated by nodes in the network. Parameters e_i and x_i have a probability interpretation: node u becomes active with respect to i, independently of its neighbors, with probability x_i. If u has $\mathbf{\Gamma}(u, i)$ neighbors that are already active with respect to i, then each one of them succeeds in activating u with probability e_i. At the end of the observation period, u becomes active with respect to i, with probability: $1 - (1 - x_i)(1 - e_i)^{\mathbf{\Gamma}(u,i)}$. Use $\mathbf{e}$ and $\mathbf{x}$ to represent the vectors of all items' endogeneity and exogeneity parameters, and use $\mathbf{\Theta} = \langle \mathbf{e}, \mathbf{x} \rangle$ to denote the vector of these pairs of values for all items.

Generative Process: Our model defines a generative process in which every item $i \in \mathcal{I}$ is given a set of chances to activate the nodes in $G = (V, E)$. Intuitively, for every item $i \in \mathcal{I}$, our model assumes *activation graph* $H_i = (V \cup \{s_i\}, E_i)$. The nodes of H_i consist of all the nodes in V plus an additional node s_i that corresponds to item i. The set of links E_i contains all the links in E plus n additional directed links $(u \rightarrow s_i)$. That is, in H_i every node *follows* the item-node s_i. Initially, only node s_i is active and the rest n nodes are inactive. An information item propagates from an active node only to its inactive followers. The activation proceeds in discrete steps. At each time step, activation of any node u, through links $(u \rightarrow s_i)$, succeeds with probability

x_i. At the same time, activation of u through links $(u \to u')$ for $u' \in V$ succeeds with probability e_i. At most one activation attempt can be made by every link. The final activation state of all nodes with respect to all items is stored in the final activation matrix $\mathbf{A}$.

Problem Definition: Given the active-neighborhood information $\mathbf{\Gamma}$ and parameters $\mathbf{\Theta}$, the likelihood of the observed activation matrix $\mathbf{A}$ can be computed as:

$$\Pr\left(\mathbf{A} \mid \mathbf{\Gamma}, \mathbf{\Theta}\right) = \prod_{i=1}^{m} \prod_{u=1}^{n} \Pr\left(\mathbf{A}(u,i) \mid \mathbf{\Gamma}(u,i), e_i, x_i\right). \tag{1}$$

Given $\mathbf{\Gamma}$ and $\mathbf{A}$, we want to estimate vectors $\mathbf{e}$ and $\mathbf{x}$ such that the compatibility between the observed activation matrix $\mathbf{A}$ and the estimated parameters, $\mathbf{\Theta} = \langle \mathbf{e}, \mathbf{x} \rangle$, is maximized. Different definitions of compatibility lead to different problems. We focus on the parameters $\mathbf{\Theta}$ that maximize the loglikelihood of the data:

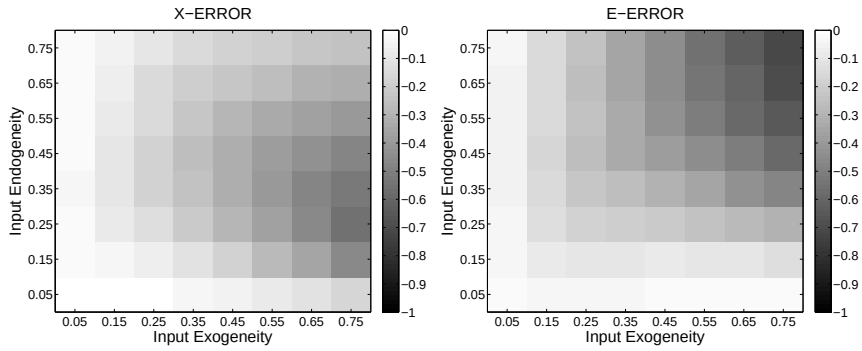
$$\mathbf{\Theta} = \arg\max_{\mathbf{\Theta}'} \mathcal{L}\left(\mathbf{A} \mid \mathbf{\Gamma}, \mathbf{\Theta}'\right) = \arg\max_{\mathbf{\Theta}'} \log \Pr\left(\mathbf{A} \mid \mathbf{\Gamma}, \mathbf{\Theta}'\right)$$

Parameter Estimation: Using Eq. (1), we rewrite the likelihood as

$$\mathcal{L}\left(\mathbf{A} \mid \mathbf{\Gamma}, \mathbf{\Theta}\right) = \sum_{i \in \mathcal{I}} \sum_{u \in V} \log\left(\Pr\left(\mathbf{A}(u,i) \mid \mathbf{\Gamma}(u,i), e_i, x_i\right)\right).$$

Thus, the parameters (e_i, x_i) of every item i can be computed independently by solving a two-variable optimization problem in the $[0,1] \times [0,1]$ range. Further, the independence of the items allows us to parallelize the item-parameter estimation. The function L_i is convex with respect to the item's parameters (e_i, x_i). Therefore, an off-the-shelf optimization method (e.g., Newton Raphson method) can be used to efficiently find the optimal values of the parameters.

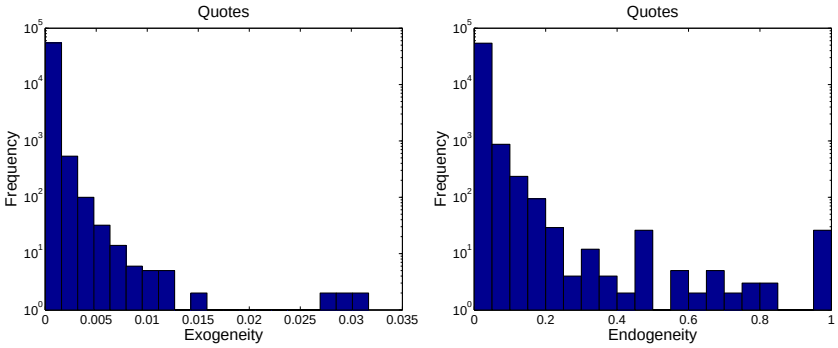
Experiments with Synthetic Data: The goal of synthetic data experiments is to study how well the parameter estimation procedure recovers exogeneity and endogeneity values. Define the *exogeneity absolute error* for the exogeneity parameters as X-ERROR$(\mathbf{\Theta}, \widehat{\mathbf{\Theta}})$ = $1/m \sum_{i \in \mathcal{I}} |x_i - \widehat{x}_i|$, where $\widehat{x}_i$ is the recovered value of the parameter x_i. The *endogeneity absolute error*, E-ERROR, is defined similarly. Figure 1 shows the results.



(a) X-ERROR (b) E-ERROR

Figure 1: Synthetic `ScaleFree` graphs: #nodes=1000, density=1%, #items=1000, endogeneity and exogeneity $\in [0, 0.8]$ (separately picked uniformly at random).

We see that the smaller the values of the input parameters, the lower the X-ERROR and the E-ERROR. Small values of these parameters generate sparse data, i.e., data with small number of activations. Real data exhibit this behavior; the most frequent item in the dataset we consider in the next section appears in less than 10% of the nodes.



(a) Exogeneity (b) Endogeneity

Figure 2: Histogram of exogeneity and endogeneity of quotes in `MemeTracker`.

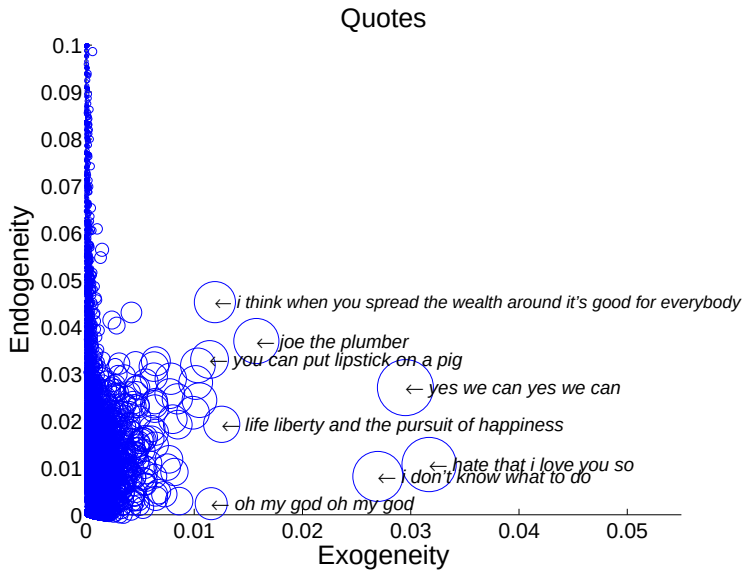


Figure 3: Scatter plot of Exogeneity and endogeneity of quotes (marker area $\propto$ frequency).

Experiments with `MemeTracker` Data: We next turn our attention to real data. We use the memetracker data available from Stanford University, which consists of quotes that have been posted on articles/blogposts from August 2008 to April 2009. Timestamps in the data capture the time that a quote was used in a post. From these data, we construct our network $G_B = (V_B, E_B)$ by selecting as nodes all the blogs hosted either by *blogspot.com* or by *wordpress.com*. For blogs $b, b' \in V_B$, there is a directed link $(b \rightarrow b')$ if there exists at least one blogpost of b linking to b'. The set of information items consists of the set of quotes that appeared in at least one blogpost of any of the blogs in V_B. We say that blog u became active with respect to quote q at time t, if t was the first timestamp that u used q in one of his blogposts.

Figure 2 plots the distribution of endogeneity and exogeneity values of the quotes. The skewed distribution of both exogeneity and endogeneity values shows that a non-negligible number of quotes are much more endogenous/ exogenous than most quotes. Figure 3 is a scatter-plot of the exogeneity and endogeneity values of the quotes. The area covered by each marker is proportional to the number of nodes it appears. For concreteness, we have also shown frequent quotes for some combinations of endogeneity and exogeneity values. Clearly, exogeneity and endogeneity are not correlated; there some quotes that have high endogeneity but low endogneity and vice versa.

Table 1: Top-5 frequent quotes.

	Exogeneity=H Endogeneity=H
1.	yes we can yes we can
2.	hate that i love you so
3.	joe the plumber
4.	i think when you spread the wealth around it's good for everybody
5.	you can put lipstick on a pig
	Exogeneity=H Endogeneity=L
1.	i don't know what to do
2.	oh my god oh my god
3.	hi how are you doing today
4.	why where are you going to john
5.	what is it
	Exogeneity=L Endogeneity=H
1.	there appears to be a sizeable number of duplicate and fraudulent applications
2.	we shouldn't let partisan politics derail what are very important things that need to get done
3.	likened zionist settlers on the west bank to osama bin laden saying both had been blinded by ideology
4.	as far as the eye can see
5.	she doesn't know yet that she has been married
	Exogeneity=L Endogeneity=L
1.	the age of turbulence adventures in a new world
2.	i've got friends in low places
3.	you shall not bear false witness against your
4.	neighbor instead of complaining about the state of the education system as we correct the same mistakes year after year i've got a better idea
5.	a woman who loves me as much as she loves anything in this world but who once confessed her...

Table 1 shows the top-5 frequent quotes for combinations of high and low exogeneity and endogeneity values. We make two observations: first, that quotes with "Exogeneity=H" exhibit shorter length than quotes with "Exogeneity=L". Second, a web search reveals that most quotes with "Endogeneity=H" were news-stories or popular quotes of the observation period. Amongst the high-exogeneity quotes, we can distinguish between those with "Endogeneity=H" and those with "Endogeneity=L". Quotes "*joe the plumber*", "*you can put lipstick on a pig*" etc. from the (H,H) bucket are front-page quotes that drew notable attention during the 2008 elections period. They are highly exogenous because they gained popularity via external media such as the television. They are also highly endogenous because they heavily propagated through the network links of the blogs. In

contrast, (H,L) quotes: "*i don't know what to do*", "*oh my god*", "*hi how are you doing today*", and "*what is it*", are popular phrases that appear in various contexts ranging from casual conversations to pop songs. Such quotes are expected to be purely exogenous – they do not trigger cascades.

Amongst the low-exogeneity quotes, we can again distinguish between those for which "Endogeneity=H" and those with "Endogeneity=L". The first correspond to long phrases that were news stories during the observation period. For example, the quote "*she doesn't know yet that she has been married*", propagated in a set of connected blogs that discussed the case of the marriage of a fourth-grade girl. Similarly, the rest of the quotes in (L,H) (except for "*as far as the eye can see*") were also news stories of that period. These are highly endogenous quotes. Compare these quotes with the quotes in bucket (L,L). Neither exogenous sources nor peer influence affect the propagation of these quotes. These are all infrequently occurring phrases, e.g., lyrics from older songs and previous year book titles.

3 Nature of People

Although E2 models the observed variation between information items, it does not capture that different people may react differently to the same information item. The E2R model incorporates a *receptivity* parameter to capture this difference in the nature of people.

E2R **Model:** Associate with every node u a parameter $r_u \in [0,1]$ that quantifies the node's tendency to be *receptive* to information items coming either from u's neighbors or from sources outside the network. Same as with e_i and x_i, r_u has a probabilistic interpretation: node u accepts any candidate activation with probability r_u. Then, the probability of the observed activation matrix $\mathbf{A}$ given the item parameters $\mathbf{\Theta}$ and user receptivities $\mathbf{r}$ is:

$$\Pr\left(\mathbf{A} \mid \mathbf{\Gamma}, \mathbf{\Theta}, \mathbf{r}\right) = \prod_{i \in \mathcal{I}, u \in V} \Pr\left(\mathbf{A}(u,i) \mid \mathbf{\Gamma}(u,i), e_i, x_i, r_u\right).$$

The probability of node u being active with respect to item i is computed as:

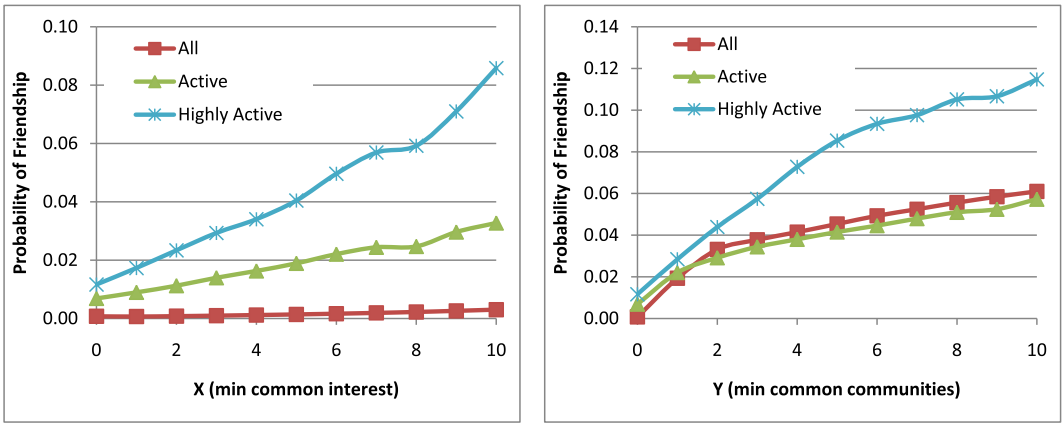
$$\Pr\left(\mathbf{A}(u,i) = 1 | \mathbf{\Gamma}(u,i), e_i, x_i, r_u\right) = 1 - (1 - r_u \cdot x_i)(1 - r_u \cdot e_i)^{\mathbf{\Gamma}(u,i)}.$$

Intuitively, every time we have an endogenous or exogenous attempt to activate a user, the user also needs to accept that activation. Receptivity is both a characteristic of the nodes and a means to allow items to reveal their true nature. Consider the extreme case of a very endogenous item that all, but a small fraction of the nodes, adopt through their neighbors. In order to capture the behavior of this minority of nodes, the E2 model would assign to i endogeneity value lower than 1. On the other hand, the E2R model will capture the behavior of these nodes through receptivity and will assign to i larger endogeneity value, allowing it to reveal its true nature.

See [1] for further details, computational techniques, and experimental results.

Dynamics of Opinion Formation: In a recent work [5], we differentiate between the innate and expressed opinions and postulate that individuals update their expressed opinions in discrete time steps by taking a convex combination of their innate opinion and the expressed opinions of their social neighbors. The weights in the convex combination depends on a user's *propensity to conform*, which is remniscent of the idea of receptivity just discussed. Through real-world experiments, they show that this value is largely specific to a given user and does not change significantly from topic to topic. In [2], we present game-theoretic models of opinion formation where opinions themselves co-evolve with friendships. In these models, nodes form their opinions by maximizing agreements with friends weighted by the strength of the relationships, which in turn depend on difference in opinion with the respective friends.

Availability, Willingness, and Knowledge: A typical person has many friends that the person can consult for opinions and advice. However, public broadcasting a question can use up social capital and the request can get



(a) Given common interests (b) Given common communities

Figure 4: Probability of friendship

lost in a myriad of status updates. Direct messaging requires manual selection and a user may have difficulty guessing which of the friends will be able to provide a quality answer in a timely manner. In [7], we describe a decision aide that provides the ranked subset of friends for a user to seek. The system mines social network data focusing on a novel set of criteria: availability, willingness and knowledge. The system response depends on (1) how likely it is that a friend is online in the near future based on past activity patterns, (2) the likelihood that a friend will respond based on the strength and nature of the interpersonal connection and past interaction behavior, and (3) a friend's knowledge and expertise on a topic and their potential for providing an informed response based on the past message content.

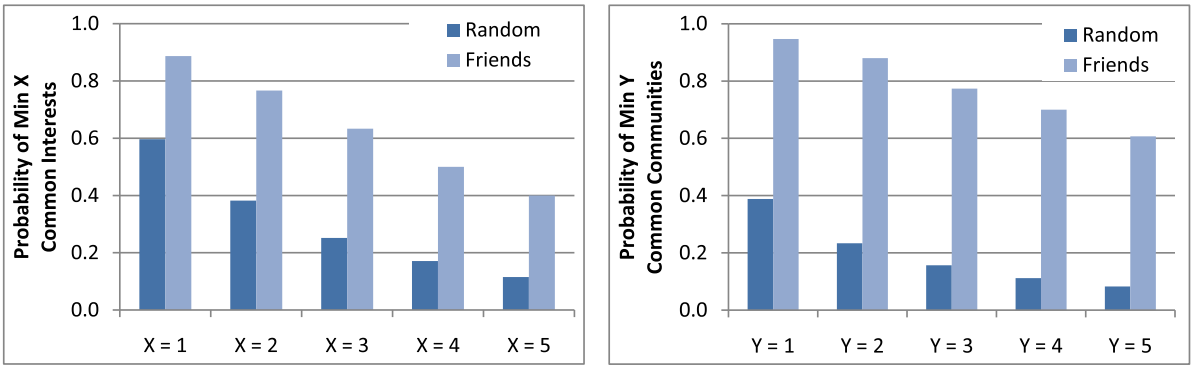
4 Nature of Relationships

Interests and Friendship: In [6], we use LiveJournal data to investigate two central questions: (1) whether having common interests makes it more likely for a pair of users to be friends, and (2) whether being friends influences the likelihood of having common interests. LiveJournal users identify each other as friends and express their interests in two ways. First, users have a list of self-proclaimed interests on their User Info page. Second, users can subscribe to communities or group blogs oriented around a given topic. We extract three binary adjacency matrices from LiveJournal data: (1) F, a user $\times$ user friendship matrix, with $F_{uu'} = 1$ iff users u and u' have friended each other, (2) I, a user $\times$ interest matrix, with $F_{ui} = 1$ iff user u specifies i as an interest, and (3) C, a user $\times$ community matrix, with $C_{uc} = 1$ iff user u watches community c.

Without any prior information, the best estimate for the probability of friendship is the fraction of random pairs that turn out to be friends. Conditional on that a pair of users share a minimum number of X interests, the probability of friendship is:

$$P(friendship \mid X) = \frac{|\{(u,u') \in U \times U \mid (\mathbf{F}_{uu'} = 1) \wedge (\mathbf{I}_u \cdot \mathbf{I}_{u'} \geq X)\}|}{|\{(u,u') \in U \times U \mid (u \neq u') \wedge (\mathbf{I}_u \cdot \mathbf{I}_{u'} \geq X)\}|},$$

where U denotes the set of users in consideration. Fig. 4(a) plots $P(friendship \mid X)$ for different values of X and different subsets of users; Active (Highly Active) users have at least ten (fifty) each of friends, interests, and communities. We see that having common interests, even just one, significantly increases the probability of friendship for all data sets. This trend is also monotonic: higher X leads to higher probability. This is a



(a) Common Interests

(b) Common Communities

Figure 5: Probability of Commonality Given Friendship

surprising outcome, given that without geographic constraint, we would not expect the conditional probability to be significantly higher. It suggests that an underlying factor is at work in LiveJournal that encourages users to make friends with those having common interests. Several LiveJournal features might contribute to this. For every interest with more than one claimant, LiveJournal provides a hyperlink to the list of users who claim that interest, thus letting one user find others to connect with on the basis of interest. Blogging and commenting is another set of activities that could help users get to know others who share similar interests.

We next investigate whether a similar relationship exists between friendship and common communities. The probability of friendship given that a user pair shares a minimum of Y common communities is:

$$P(friendship \mid Y) = \frac{|\{(u,u') \in U \times U \mid (\mathbf{F}_{uu'} = 1) \wedge (\mathbf{C}_u \cdot \mathbf{C}_{u'} \geq Y)\}|}{|\{(u,u') \in U \times U \mid (u \neq u') \wedge (\mathbf{C}_u \cdot \mathbf{C}_{u'} \geq Y)\}|}.$$

Fig. 4(b) plots $P(friendship \mid Y)$ for different Y values and data sets. We observe similar trends as those in Fig. 4(a): a user pair is monotonically more likely to consist of friends if they share more common communities.

To study the second question raised at the beginning of this section, we write the probability that a pair of friends shares at least X common interests as:

$$P(X \mid friendship) = \frac{|\{(u,u') \in U \times U \mid (\mathbf{F}_{uu'} = 1) \wedge (\mathbf{I}_u \cdot \mathbf{I}_{u'}) \geq X\}|}{\sum_{u \in U} \sum_{u' \in U} F_{uu'}}.$$

Fig. 5(a) compares $P(X \mid friendship)$ to $P(X)$ for different values of X on Highly Active subset of users. Similar trends are observed on other datasets. It shows that for every X, $P(X \mid friendship)$ is significantly higher – between 1.5 and 3.5 times higher – than $P(X)$. The likelihood of common interests conditioned on friendship is as high as $P(X = 1 \mid friendship) = 0.89$ and $P(X = 2 \mid friendship) = 0.77$. This result suggests that friendship is a potentially significant source of signals in inferring a person's interests.

We conducted a similar exercise on communities. Fig 5(b) plots $P(Y)$ and $P(Y \mid friendship)$ for various Y's and for the Highly Active dataset. We see similar trends as in Fig. 5(a), but the difference is even higher. $P(Y \mid friendship)$ is 2.4 to 7.3 times higher than $P(Y)$, suggesting that friendship is an even stronger signal in detecting common communities.

See [6] for extensions where friendship has strength associated with it.

Group Participation: In [3], we study what makes a person become a member of a group. We addressed this question in the context of Twitter chats, which are time-bound synchronous group interactions carried out in

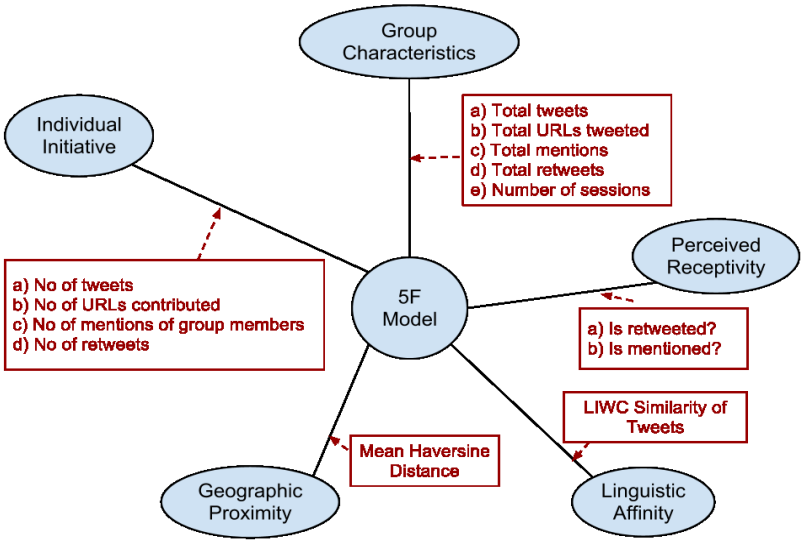


Figure 6: Overview of the *5F Model*

real time on a focused topic. For instance, #engchat is a chat about English education held at 7-8pm EST on every Monday. During a chat session, the participants continuously interact on the designated topic by tweeting their opinions and marking their tweets with the hashtag of the particular chat group. While weekly groups like #engchat are the most common ones, there are others such as #mathchat that meet twice a week, #collegechat that meet bi-weekly or #edchat that are week-long conversations. Most of the chat groups also have dedicated blogs that provide various resources such as transcripts of past sessions and schedule of upcoming discussions. In a companion work [4], we describe algorithms for mining chat groups from Twitter data stream.

We developed *5F Model* that predicts whether a person attending her first chat session in a particular Twitter chat group will return to the group. This model, pictorially depicted in Figure 6, considers five different classes of factors: *individual-initiative, group characteristics, perceived receptivity, linguistic affinity* and *geographical proximity*. For example, the number of tweets, the number of URLS in the tweets, the number of mentions and retweets contributed by the person during her first session provide indication of her individual initiative. Using data from thirty education-related chat groups, we study the predictive power of these factors individually as well as collectively. We use logistic regression for statistical analysis and a *Pseudo-R* measure (Nagelkerke R^2 Index) to compare the models.

The regression results are summarized in Table 2. This table has four columns. The first column is the name of the model and corresponds to one of the five factors. The second column lists the Twitter specific variables used for each of the corresponding factors. The third column consists of two subcolumns. The first subcolumn shows the cofficients of the corresponding explanatory variables in the individual-level models, whereas the second subcolumn gives the coefficients for the unfied model. The third column gives the pseudo-R measure for the individual models. The pseudo-R value for the unfied model is 0.14 and is shown at the bottom of the table. The statistically significant variables are marked with * for p-value < 0.05, ** for p-value < 0.01 and *** for p-value < 0.001.

Individual initiative model: The results show that all the variables except for *usermentions* are statistically significant. The number of tweets are positively correlated with returning to the chat group, emphasizing the predictive power of early interest exhibited by the user. The variable *userurl* is negatively correlated with returning to the group. One possible explanation for this result can be given as follows: For users that share a large number of urls, i.e. users that already acquire a certain level of knowledge, the added informational gain

Factors	Variables	Coefficients		Pseudo-R
		Individual Model	Unified 5F Model	
Individual Initiative	usermentions	-0.016	-0.007	0.09
	userretweets	-0.13***	-0.077***	
	userurl	-0.16***	-0.092***	
	usertweetcount	0.147***	0.05***	
Group Characteristics	groupmentions	-0.0001	-0.0004	0.03
	groupretweets	0.0014*	0.002***	
	sessionurl	-0.003***	-0.002*	
	sessiontweetcount	-0.0005	-0.0008*	
	groupmaturity	-0.01***	-0.007***	
Perceived Receptivity	ismentioned	1***	0.445***	0.08
	isretweeted	0.69***	0.24	
Linguistic Affinity	liwccors	2.159***	1.215***	0.1
Geographical Proximity	distance	-0.00005***	-	0.01

Pseudo-R for the unified *5F Model* = 0.14

* $p < .05$, ** $p < .01$, *** $p < .001$

Table 2: Results of Statistical Analysis

from chat sessions can be smaller, resulting in less incentive to attend future sessions.

The negative correlation for *userretweets* indicates that retweeting behavior can be used to distinguish *real* participants of chat groups from those that are merely retweeting the tweets of their friends who are attending a chat session. Consider the following illustrative scenario. Assume that *user1* attending *#1stchat* shares a tweet "Check out article bit.ly/342dfser #1stchat". This tweet is seen not only by the attendees of *#1stchat* but also the followers of *user1*. One such follower, say *user2*, can find the tweet interesting and retweet it. Here, *user2* who *appears* to be attending his first *#1stchat* session may not return to this group.

Group characteristics model: Statistically significant variables are *groupretweets, sessiontweetcount, sessionurl* and *groupmaturity*. Capturing the significance of information overload, *sessionurl* and *sessiontweetcount* have negative correlation. The variable *groupmaturity* has negative correlation with the odds of come back, i.e. users that attempt to join more mature groups are less likely to return to the group. The results also indicate the significance of *informational influence* as demonstrated by the statistical significance and positive correlation of *groupretweets*. However we observe that the correlations of these factors are relatively mild. For instance, an increase of 1 retweet in group discussion decreases the log odds of come back by 0.0014. *Pseudo-R*(=0.03) values for this model are worse when compared to those of *individual initiative model*, showing that *individual initiative* factors are relatively better indicators of future participation.

Perceived receptivity model: Our results show the importance of social inclusion in ensuring continued participation. For instance, the log odds of returning to a group increases by 1 if a user is mentioned in the first session that he/she attends. Similarly, the odds of returning improves by 0.69 if the user is retweeted by others in the chat session. Both of these findings are statistically significant. This result is in agreement with relevant research in other online communities.

Linguistic affinity model: We make use of the *Linguistic Inquiry and Word Count (LIWC)* tool to compare linguistic markers between a user and a group. We consider the set of tweets a user shares in her first session as a text document and compute the value of each linguistic marker to obtain her *LIWC-vector* for that particular session. Similarly, we aggregate all the tweets from other users and compute the *LIWC vector* of the group. To compute affinity, we use Pearson correlation measure. We find that linguistic affinity is statistically significant

and highly correlated with returning to a chat group, which is in line with research in social sciences, particularly the *speech codes theory*. The highest *Pseudo-R* value for this model shows that the linguistic characteristics are the best indicators of future participation.

Geographical proximity model: To study the influence of geographical proximity, we calculate the mean distance of the user to everyone else in the group using the Haversine formula. The location for each user is determined based on the location field of the user profile. We see that returning to a group is only mildly correlated with geographical proximity. An increased distance of 1km reduces the log odds of returning to the group by only 0.00005. Regression tasks performed *per-chat group* showed that geographical proximity is statistically significant for only seven educational Twitter chats. Two chats had positive correlation and five had negative correlation. For instance, #globalclassroom has positive correlation with the variable *distance*, indicating the positive effect of diverse locations in returning to the group. Such behavior is to be expected given the global goal of this particular group. Yet groups like #jedchat have negative correlation with increased distance. This group is on Jewish education and is mostly popular in Israel. Overall, the *Pseudo-R* value for this model is the worst among all models, showing that geographical characteristics are generally poor indicators of future participation.

Unified 5F Model: In this model, we consider all the explanatory variables in conjunction, except geographic proximity (*distance*). The reason for omitting the latter is that we could determine the location of only a subset of users and this factor anyway turned out to have limited fit. As expected, this model has the largest *Pseudo-R* value. Each independent variable has similar explanatory trend as we observed with individual models.

User Survey: We complemented the results from the statistical data analysis with a user survey to directly understand from users involved in Twitter chats their attitudes towards these chats. The survey had three main parts, addressing questions related to: (1) usage, advantages and disadvantages, (2) sense of community and responsibility, and (3) evolution of participation. The survey was tweeted through the hashtag of each chat group studied. Respondents of the survey were encouraged to share the survey with their Twitter followers.

The survey results highlighted various distinctions between Twitter chats and other online groups and face-to-face discussions. We found *informational* support to be more important to Twitter chat members than *emotional support*. Although prior work suggests that informational support is negatively correlated with the sense of community, we found the sense of community to be very strong in Twitter chats. In fact, its members communicate with one another outside chat sessions much more than expected from the literature. Disadvantages identified by the survey respondents also mark an interesting distinction between Twitter chats and other online groups. While for other online communities, the lack of face-to-face interactions is a main disadvantage, Twitter chat users focus on the content. More specifically, due to the synchronous and open nature of Twitter, the pace of information is the biggest challenge of Twitter chats.

The survey results reinforced most findings of the statistical analysis. Groups becoming closed to new members over time (as captured by *groupmaturity* in our model) is seen anecdotally in survey results. The importance of social inclusion is also observed in the responses of two survey participants that reduced (one ending) their participation due to the lack of receptivity. The geographical diversity listed as an advantage in the survey also indicates that geography is not a limiting factor for Twitter chats.

5 Ongoing Work

Social analytics continue to provide us opportunities for exploring critical issues and building useful systems. Some of our current research directions include:

- Many users of social media entertain an illusory sense of "privacy by hiding in the crowd". We are interested in ascertaining if one could accurately determine a user's attributes by building an inference

system over the history of user interactions, and thus shattering this illusion. We are also interested in exploring what a user can do in order to achieve privacy (short of not participating in social media).

- A huge potential exists to leverage aggregate information from social media, news sites, and the internet as a whole for enterprise and market insights as well as enabling interesting user applications. We aim to ingest, mine, and analyze such information in order to enable a wide variety of social intelligence applications and provide useful insights by identifying interesting patterns, alerting users to unusual or trending events, allowing adhoc, "what if" analysis, and other capabilities.

References

[1] R. Agrawal, M. Potamias, and E. Terzi. Learning the nature of information in social networks. In *ICWSM*, 2012.

[2] K. Bhawalkar, S. Gollapudi, and K. Munagala. Coevolutionary opinion formation games. In *STOC*, pages 41–50, 2013.

[3] C. Budak and R. Agrawal. On participation in group chats on twitter. In *Proceedings of the 22nd international conference on World Wide Web*, pages 165–176. International World Wide Web Conferences Steering Committee, 2013.

[4] J. Cook, K. Kenthapadi, and N. Mishra. Group chats on twitter. In *Proceedings of the 22nd international conference on World Wide Web*, pages 225–236. International World Wide Web Conferences Steering Committee, 2013.

[5] A. Das, S. Gollapudi, R. Panigrahy, and M. Salek. Debiasing social wisdom. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 500–508. ACM, 2013.

[6] H. Lauw, J. C. Shafer, R. Agrawal, and A. Ntoulas. Homophily in the digital world: A livejournal case study. *Internet Computing, IEEE*, 14(2):15–23, 2010.

[7] A. Nandi, S. Paparizos, J. C. Shafer, and R. Agrawal. With a little help from my friends. In *Proceedings of the 2013 IEEE International Conference on Data Engineering (ICDE 2013)*, pages 1288–1291. IEEE Computer Society, 2013.