On the Precision of Social and Information Networks

Reza Bosagh Zadeh
Stanford University
rezab@stanford.edu

Ashish Goel
Stanford University
ashishg@stanford.edu

Kamesh Munagala
Duke University
kamesh@cs.duke.edu

Aneesh Sharma
Twitter, Inc
aneesh@twitter.com

ABSTRACT

The diffusion of information on online social and information networks has been a popular topic of study in recent years, but attention has typically focused on speed of dissemination and recall (i.e. the fraction of users getting a piece of information). In this paper, we study the complementary notion of the *precision* of information diffusion. Our model of information dissemination is "broadcast-based", *i.e.*, one where every message (original or forwarded) from a user goes to a fixed set of recipients, often called the user's "friends" or "followers", as in Facebook and Twitter. The precision of the diffusion process is then defined as the fraction of received messages that a user finds interesting.

On first glance, it seems that broadcast-based information diffusion is a "blunt" targeting mechanism, and must necessarily suffer from low precision. Somewhat surprisingly, we present preliminary experimental and analytical evidence to the contrary: it is possible to simultaneously have high precision (i.e. is bounded below by a constant), high recall, and low diameter!

We start by presenting a set of conditions on the structure of user interests, and analytically show the necessity of each of these conditions for obtaining high precision. We also present preliminary experimental evidence from Twitter verifying that these conditions are satisfied. We then prove that the Kronecker-graph based generative model of Leskovec *et al.* satisfies these conditions given an appropriate and natural definition of user interests. Further, we show that this model also has high precision, high recall, and low diameter. We finally present preliminary experimental evidence showing Twitter has high precision, validating our conclusion. This is perhaps a first step towards a formal understanding of the immense popularity of online social networks as an information dissemination mechanism.

Categories and Subject Descriptors

H.1 [**Information Systems**]: Models and Principles

COSN'13, October 7–8, 2013, Boston, Massachusetts, USA.

http://dx.doi.org/10.1145/2512938.2512955.

Keywords

Social networks; Precision; Recall; Modeling

1. INTRODUCTION

Modern social and information networks such as Facebook, LinkedIn and Twitter are used by hundreds of millions of users every day. There are many hypotheses as to the source of their popularity, and one popular hypothesis relates to the effectiveness of these networks as information dissemination mechanisms [10, 28]. In particular, a fundamental question about effectiveness is one of personalization: given the large number of users, one would expect them to be interested in a diverse set of content, and the network must be an effective information conduit, simultaneously, for all of them. Given that information dissemination mechanism in these networks occurs via *broadcast* (as opposed to pairwise interactions) over the network topology, it is apriori unclear whether effective information dissemination is even feasible. For instance, wouldn't users receive a large amount of un-interesting content via this mechanism? And complementarily, wouldn't users miss a large amount of content they would have potentially been interested in?

The starting point of our study is this commonly stated belief, especially in the media, that online social and information networks mostly generate information that is irrelevant for most users [24]. This claim is often based on inspecting a random tweet. However, such a claim ignores the interest-based construction of social networks: as suggested earlier, users on any social or information network have diverse interests, and tend to *follow* (i.e., receive content from) other users who share some of their interests and post content that is interesting to them. Thus, although a random tweet on Twitter is uninteresting to a random user, it could be that for any given user, the tweets in their timeline are very relevant to them.

The study of usefulness (or relevance) of content has been a primary theme in the information retrieval literature [22], but to the best of our knowledge it has not been directly studied for information diffusion on social and information networks. We adapt the widely accepted definition of relevance for networks by defining the *precision* of information in a social network: the fraction of content received by a user that is relevant to them, where relevance is captured by a match between the content and some "interest" of the user. Then we capture virality by defining two quantities: The *recall*, which is the fraction of content relevant to a user that (s)he does receive, and the *dissemination time* for content, *i.e.*, the number of hops in the graph taken for this

content to spread to users who would be interested in this information.

Assuming that users have interests, and the social network is constructed according to users' interests, the following natural questions arise about the precision and recall: *What conditions (if any) on the structure of user interests are necessary for a social and information network to ensure users have high precision and recall, and dissemination time is small? Can we empirically validate these conditions as well as the conclusion on existing networks?*

We motivate this question with a preliminary empirical user study that attempts to directly measure relevance without resorting to a definition of user interests: we ask 10 active Twitter users to rate a set of 30 tweets as Relevant/Not Relevant. The users are students at Stanford University who log in at least once a week on average, follow at least 30 people, and receive at least 20 new tweets a week in their timeline. The set of 30 tweets is put together by choosing 15 tweets from the user's timeline in the past 7 days, and 15 unique randomly selected tweets out of the set of all tweet impressions (or tweet renderings) over the same 7 days[1]. The set of 30 tweets is then rendered in a random order as per usual tweet rendering guidelines [11]. The *precision* of each set of 15 tweets is then the fraction of these tweets that the user marks as being relevant. The results of the experiment for each of the 10 users is shown in Figure 1. The average precision of users for tweets drawn from their timeline is 70%. On the other hand, the precision drops to around 7% for the set of random tweets shown to the users! Even though this is too small a user study to draw a definitive conclusion about the actual value of precision on Twitter, the results lend some credence to the hypothesis that social networks such as Twitter are much more precise than one would expect if users were seeing content at random. Note that since we showed (as control) each user 15 random tweets chosen from tweet *impressions*, and got a low relevance score for this control set, it does not appear that inspection paradox[2] alone could be an adequate explanation of the high precision we see in this trial.

1.1 Necessary Conditions for Precision

In this paper, we first outline some necessary conditions for obtaining high precision. For each of these conditions, we state the hypothesis, validate it with data, and argue via modeling and analysis, why the hypothesis is necessary for obtaining high precision.

Interest-based Networks.

Our first hypothesis is a natural one: Users on social and information networks have interests, and link to other users who share some or all of these interests. This assumption is folklore in how these networks are generated –several commonly used generative models of social networks indeed use this assumption [18, 17, 7]. We define (in Section 2) an analytic model capturing the essence of these generative models: There are a set of users V and a set of interests I. Each user $u \in V$ has a set of interests $C(u)$ that (s)he is interested in. We term these users *consumers* for interest i. Each user connects to other users based on their interests, and this yields a graph $G(V, E)$ on the users, which is the observed social network. This network could be directed (e.g., Twitter), where some users follow others and information flows along directed edges, or undirected (e.g., Facebook), where friendship is mutual, and information can flow in both directions along an edge.

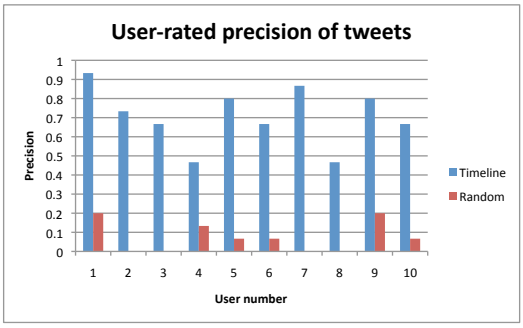

Figure 1: Comparison of self-reported precision between tweets from a user's timeline and tweets chosen at random.

In order to analyze precision in this model, we need to define which users sharing an interest $i \in I$ *produce* content related to the interest. Let $P(i)$ denote the set of users who act as producers, We show (in Section 3) that if for all interests i, $P(i) = C(i)$, which means any consumer can be a potential producer, then it is only possible to construct networks with good precision in the trivial scenario where all users have the same interests.

Production vs. Consumption.

This leads us to our second hypothesis: the production interests of a user are narrower than the consumption interests. In other words, $P(i) \subset C(i)$. We validate this assumption on Twitter (described in Section 2). We define production as either tweeting or retweeting a tweet, and consumption as clicks by the user on tweets that contain a URL. For simplicity, we refer to this as a click on a tweet. We show that the set of interests captured by clicks has larger entropy (per user) than the set capturing tweets or retweets. We note that both restricting attention only to tweets containing URLs, and requiring clicks as a measure of consumption interests are strict notions, which makes the empirical results stronger.

We also show via analysis (in Section 3) that separation of production from consumption is still insufficient to explain high precision. In particular, we show that if users choose their production and consumption interests at random from any distribution over interests (subject to mild restrictions), it is not possible to achieve even constant precision. Our result is fairly robust to the empirically observed variability in the number of user interests, and the cardinality of the interests. In Appendix A, we show the same result when users themselves have varying number of interests, as in the affiliation network models [17, 7].

[1]We imposed two restrictions on the randomly selected tweets: the tweets must be in english (all the survey takers were english speakers), and the tweet must not be a reply (since a reply may not make sense outside of the full conversation, thus yielding artificially low precision).

[2]The inspection paradox is an analogue to the well-known friendship paradox [6]: high quality users have more followers and hence a random tweet impression is of higher quality than a random tweet.

Structured Interests.

The above result makes a case for interests with structure: Users do not choose interests randomly, but rather, choose them in a correlated fashion. In other words, interests have a correlation structure, and users are more likely to choose from among correlated interests than from among uncorrelated interests. We verify this assumption by measuring the correlation between interests on Twitter defined by the overlap between the sets of users having these interests. We show that the correlation is indeed much larger than what can be expected had users chosen interests at random. We then cluster the interests using these correlations, and show that these clusters have natural interpretations – sports, art, technology, etc.

It would therefore appear that users are defined by their values on various *attributes* (sports, art, etc), and interests themselves are defined either as these attributes or sets of attributes taking specific values. We finally consider a generative model of social networks that is based on users having attributes: This is the Kronecker graph model [18, 14, 21], where users connect with other users based on similarity in attribute values. We define interests using the attributes in this model, as well as producers and consumers of these interests in a natural way, so that producers are more aligned with an interest in terms of attribute similarity than consumers. We show (in Section 4) that the resulting user-user graph (or social network) has perfect precision and recall, and constant dissemination diameter for any interest.

Finally, we present (in Section 5) an empirical study to measure the precision of Twitter, defined as the fraction of the set of interests that a user receives from her friends that she is actually interested in consuming. As before, we use clicks on URLs within tweets as a proxy for consumption interest. We observe an average precision of 40%. This implies on average, users are interested in one in 2.5 topics (or interests) their neighbors tweet about. While this is already a surprisingly good number, it is worth repeating that clicks on URLs in tweets (and restricting attention to tweets with URLs) are a strict notion for capturing user interests, and it is conceivable that we are under-estimating precision in our experiments. For all our experiments, we use a classifier trained within Twitter to assign topics or interests to a tweet.

In summary, we show, both by theoretical as well as empirical analysis, that it is indeed possible for a social network to have high precision and recall for broadcast information dissemination if (a) users have interests, and connect with other users based on similarity in interests; (b) the producers of an interest are a small subset of consumers; and (c) users don't choose interests at random, but the interests have structure defined by attributes, which also define the users. We consider this to be a surprising result since a priori, a low dissemination time seems to require a well connected network which seems to trade-off with precision (this is analogous to the well studied precision-recall trade-off in Information Retrieval community).

Caveats.

We should emphasize that our results should be viewed as a first step in the understanding of the theoretical and empirical underpinnings of precision in information dissemination. For instance, our empirical measure of precision is somewhat primitive (based on broad interests) and can be refined. Though we have made use of proprietary click data in our empirical analysis of precision, we believe our user study provides an empirically better and reproducible template for measuring precision across social networks, and as future work, we plan to replicate it on a larger scale in a more principled fashion. In summary, each of our hypotheses presented above is a valid area of research in itself, deserving a more in-depth study with fine-tuned metrics, experiments, and theory. We discuss future research directions in Section 6.

1.2 Related Work

The rise of the World Wide Web and online social networks has seen an explosion of interest in the structure of these networks. In particular, researchers have made many empirical observations about network structure and posited models that could explain this structure. A comprehensive survey of these is beyond the scope of this work, but we mention some relevant works (and surveys/books, where available). There is a long line of work studying the power-law degree distributions that arise in networks [26, 3, 5]. Among other structural properties that have been studied extensively are small diameter [25, 31, 5], navigability [16, 5], densification [20] and clustering coefficients [8, 5]. It is important to note that much of the above work not only identifies the relevant structural properties, but also proposes models of phenomenon that could give rise to those properties. Among the desiderata for such models is mathematical tractability and statistical soundness in that it's assumptions and predictions match well with empirical data.

Since the focus of this paper is the interplay between network structure and information dissemination, we focus our attention to modeling approaches that seek to explain properties related to information dissemination. The empirical study of information dissemination through social networks, and the role of network structure in this process, has a long history in sociology [9, 29]. There has been an explosion of work from the computer science community in this area (sometimes known as viral marketing [19]) since the influential works of [4, 13], and we refer the interested reader to a slightly old but excellent survey of work in this area [15].

Another line of work in network modeling is relevant to us, namely one that seeks to capture the role of *user interests* in the formation of the network. The works we are aware of are the Kronecker graph model [18], the MAG model [14], the affiliation networks model [17], and a network model based on user behavior [7]. Among these, both the Kronecker graph model and the MAG model seek to be both mathematically tractable and statistically sound. On the other hand, the affiliation networks model and the model based on user behavior are theoretical.

We note that many of these models study the role of network structure in information propagation, but to the best of our knowledge none of them have studied the trade-off between precision and recall. Recall in the broadcast model has been extensively studied in the context of rumor spreading, but the goal in that line of work has typically been to maximize speed of propagation [12]. We are not aware of any work studying precision of information in social networks.

Finally, we mention that precision and recall are extremely well-studied concepts in the Information Retrieval community [22]. In particular, they are arguably the two most

frequent and basic measures for information retrieval effectiveness.

2. A USER INTEREST MODEL

We start by formally describing the framework in which we analyze the precision of information diffusion. We describe a general model, folding into it the first two hypotheses presented above – that users have interests, these interests determine who they connect to, and that for any user, the set of interests for which they are the producers is a subset of the set of interests they consume (or are interested in). After presenting the model, we present an empirical rationale for the basic premise of our model.

2.1 An Interest Based Model For Precision and Recall

Interest Graph.

The set of users in the social network is denoted by set V. Every user $u \in V$ is assumed to have a fixed set of *interests*. For user u, we denote the set of interests by $C(u)$. This defines a natural *user-interest* graph $Q(V, I, F)$, where I is the set of interests, and F is the set of user-interest edges in this graph. In the discussion below, unless otherwise stated, we use $n = |V|$ and $m = |I|$.

The interests themselves are defined via the set of users that have those interests: Each interest $i \in I$ is defined by a set of *producers* $P_i \subseteq V$ and set of *consumers* $C_i \subseteq V$, such that $P_i \subseteq C_i$. The set C_i is precisely the set of neighbors of i in the graph Q, and captures all nodes interested in reading/consuming content related to interest i. The producers are a subset of these users that are sufficiently interested to produce new content or rebroadcast content associated with interest i.

We say that $Q(V, I, F)$ is *undirected* if for any interest $i \in I$, the set $P_i = C_i$, *i.e.*, each consumer of i is a potential producer as well. If this condition is not true so that $P_i \subset C_i$, we call the graph Q as directed. We present the rationale for the separation between producers and consumers below.

User Graph.

As is customary in the literature, we represent the social network as a directed *user-user* graph $G(V, E)$. In such a network, if there is a directed edge from u to v, then we assume u *follows* v, and information broadcast by v is received by u. We call u a *follower* of v. (A user-user graph $G(V, E)$ is undirected if for any $(u, v) \in E$, u follows v and v follows u. An example of such a network is Facebook, where friendships are undirected.) This user-user graph $G(V, E)$ is constructed by the users based on the structure of the user-interest graph Q, *i.e.*, users form links based on mutual interests in some specific manner. At the very least, for any edge $(u, v) \in E$, some interest of user u must be the same as some interest of user v. The nature of this link generation process (in addition to the structure of the interests) will be critical to how information disseminates in G.

DEFINITION 2.1. *Given a directed user-user graph $G(V, E)$, define the following quantities: Let $N(u)$ denote the set of nodes that u* follows. *Let $P(v) = \{i | v \in P_i\}$ and let $C(v) = \{i | v \in C_i\}$. Finally, let $S(v) = \{i | \exists (u \in N(v) \wedge i \in P(u))\}$,* i.e. *the size of the union of the production interests of the users who v follows.*

The above definition can be extended to analogous terms for undirected graphs $G(V, E)$.

Information Dissemination Metrics.

An *event* refers to a piece of information that corresponds to a single interest i, and originates at one user $v \in P_i$. This information proceeds along the edges of the social network according to the following *broadcast* process: At any time t, suppose the event has been received by a set R_t of nodes; initially, $R_0 = \{v\}$. Let $Q_t = R_t \cap P_i$ denote the nodes in R_t which are producers. These nodes broadcast the event to their followers, and the set R_{t+1} is updated by including these followers. The process terminates when the set of receivers does not increase from one step the next. Let $R_i(v)$ denote the final set of receivers if the broadcast started at node $v \in P_i$. Our model of propagation is rather simplistic, and it would be interesting to expand our results to models where resending a piece of information is based on a stochastic process or the "importance" of the information (eg. [13]).

Our goal is to study what user-interest graphs and what generative processes of user-user graphs lead to "good" information dissemination. We will make the following simplifying assumption: *The user-user graph enforces that all producers P_i of an interest i are strongly connected, so that they can both send as well as receive information related to interest i.* We capture the quality of the information dissemination via the following metrics.

DEFINITION 2.2. *Given a user-interest graph $Q(V, I, F)$ and associated user-user graph $G(V, E)$, the* precision *of a user v is defined as $\frac{|C(v) \cap S(v)|}{|S(v)|}$. This measures the fraction of interests that v receives that it is actually interested in. The* recall *of a user v is defined as $\frac{|C(v) \cap S(v)|}{|C(v)|}$. This measures the fraction of v's interests that it actually receives.*

We consolidate the above two measures into the following notion of α-PR user-interest graphs.

DEFINITION 2.3. *A user-interest graph $Q(V, I, F)$ is said to be α-PR if there exists a user-user graph $G(V, E)$ such that:*

$$\min_{v \in V} \frac{|C(v) \cap S(v)|}{|C(v) \cup S(v)|} \geq \alpha$$

Analogously, a user $v \in V$ is said to be α-PR if

$$\frac{|C(v) \cap S(v)|}{|C(v) \cup S(v)|} \geq \alpha$$

DEFINITION 2.4. *The* dissemination time *of the event is the number of iterations of the broadcast process before the event reaches all nodes in C_i.*[3]

The main question we ask can now be phrased formally as follows: *What kind of user-interest graphs and user-user graphs based on these interests lead to high precision and recall (captured by α-PR for constant α) and constant diameter in the above broadcast process? And is there a generative process that would allow emergence of such graphs?* An important special case of interest is the following:

[3]In all the models we consider, the graph is sufficiently connected that the event reaches all nodes in C_i with high probability.

DEFINITION 2.5. *A user interest-graph* $\mathbb{Q}$ *is* PR-perfect *if it is* α*-PR for* $\alpha = 1$.

PR-perfectness of a user-interest graph means that there is an associated user-user graph where information dissemination has 100% precision and recall, *i.e.*, all pieces of information a user receives are relevant, and furthermore, the user receives all relevant information.

2.2 Empirical Validation of Production vs Consumption Interests

A basic premise of the model described above is that users have distinct consumer and producer interests. We validate this premise empirically by using data from Twitter. As a side-effect of this analysis, we demonstrate that the production interests are in fact substantially "narrower" (i.e. have smaller entropy) than consumption interests, which plays an important role in subsequent analysis.

Experimental Setup.

We use a classifier trained within Twitter Inc. that can tag the content of a tweet with topics (which we interpret as interests for the purpose of this paper). For our classifier we used L_2 regularized Multinomial Logistic Regression trained with stochastic gradient descent over a training corpus where the number of examples for the 48 classes considered ranged from 5K to 30K. To classify tweet content we converted the text to lower case, removed embedded urls, if any, and represented each the content as a bag of character 4-grams. The set of unique feature IDs was hashed onto a 1M dimensional space but no feature selection was performed. While we have not tuned this model very extensively, it performed adequately and on par with other representations (e.g., tf-idf weighted unigrams). The training instances were collected via combination of manual labeling and manually constructed heuristic rules transferring labels from specific authors, urls, or hashtags. While we used a custom learner implementation, very similar results can be obtained with open source tools, such as Mahout, Mallet or sofia-ml [2, 23, 30].

Note that the classifier only uses features from the text of the tweet, guaranteeing that the topics tagged do not use the social network (this is going to be important later when we use the same classifier to acquire a lower bound on precision). The mean AUC (Area Under the receiver operating characteristic Curve) across the set of topics is 0.914, ranging from 0.97 down to 0.80, suggesting that the classifier is high quality. The classifier provides 48 topics, which are listed in Figure 4. The entropy for distributions over interests ranges from 0 to $\log(48) = 3.87$.

Empirical Analysis.

We now present preliminary empirical evidence that Twitter users have narrower production interests than consumption interests. As before, we generate production and consumption interests for users in the following manner: we obtain the set of production interests for a user by taking all the tweets (including retweets) produced by a user and tagging each tweet with topics with the same classifier as the one used in Section 2.2. For the consumption interests, we again resort to looking at tweets where the user explicitly expressed an interest in the tweet via clicking an URL in the tweet. Note that in order to do this, we restrict attention only to tweets that contain an URL. To be clear, this encompasses tweets containing pictures, videos etc since their representation in tweets is via an URL. We emphasize that our definition of consumption is narrow both due to the filtered selection of tweets with URLs and also due to the fact that a click can be construed as a more definite indication of interest, as opposed to simply receiving the message. Hence, we expect (though we have not formally proved it) that the true consumption interests are wider than suggested by this study, and would further widen the separation between production and consumption interests.

We again generate a user sample of interest as before: we compute PageRank [27] on the follow graph, then from the 10 million highest PageRank users, we uniformly sample 1000 users *who have generated at least* $k = 10$ *tweets and clicks in a given 10 day period.* This allows us to avoid using dormant users and spammers in the analysis, and ensures we have enough tweets to analyze the production and consumption distribution. We then tag k uniform tweets that a user clicked on in their timeline in the given interval. The tags from the classifier give us a probability distribution over topics for the consumption of the user, with a corresponding entropy. Similarly, we tag k uniform random tweets that the user produced, giving a production distribution with corresponding entropy.

Distribution	Average Support	Average Entropy
Consumption Interest	7.78	1.999
Production Interest	3.96	1.242

Figure 2: The distribution of production vs consumption interests

Our results are summarized in figure 2. It is clear that in terms of both the support over interests and the entropy, the distribution of consumption interests is much broader than production interests. The average support and entropy are obtained by averaging the support/entropy of the production/consumption distribution over all users. We get similar quantitative and qualitative results when we vary the time period and k.

3. NECESSARY CONDITIONS ON THE USER-INTEREST GRAPH

Our goal in the next two sections is to understand whether it is possible to have non-trivial PR-perfect User-Interest graphs. Towards this end, we will now develop two necessary conditions that such graphs must satisfy. In the previous section, we already presented empirical evidence that users have narrower production interests than consumption interests. In this section, we first prove that if for every user, her production and consumption interests are identical, then the corresponding User-Interest graph can not be PR-perfect.

We then show that a user-interest graph that is formed by users choosing production and consumption interests uniformly at random from a distribution over interests cannot be constant-PR. This result is quite robust: In the full paper, we extend it to random graphs where the users also have non-uniform degrees. We will then empirically examine the user-interest graph of a subset of Twitter, and exhibit non-

trivial structure suggesting that this graph is not drawn from a random graph model with a given degree distribution.

3.1 Production vs. Consumption

We prove that it is not possible to achieve PR-perfection with non-trivial undirected user-interest graphs (i.e. with consumption and production interests being identical).

LEMMA 3.1. *If both the user-user graph $G(V, E)$ and user-interest graph $Q(V, I, F)$ are undirected, and if $G(V, E)$ is connected, then the only PR-perfect graph is a graph where every user has identical interests.*

PROOF. If $(u, v) \in E$, then for any i such that $u \in P_i$, v receives information related to i from u. If G is PR-perfect, then $v \in C_i = P_i$. Since the graph is connected, all nodes must share the same interests. □

LEMMA 3.2. *Suppose $Q(V, I, F)$ is undirected, while $G(V, E)$ is directed. If Q is PR-perfect, then for any strongly connected component $S \subseteq V$ of G, all users in S share the same interests.*

PROOF. Suppose $v \in P_i$. Then for edge $(u, v) \in E$ so that u follows v, it must be that $v \in C_i = P_i$. Therefore, all users in a strongly connected component must have the same interests. □

The above two observations justify making Q directed, *i.e.* assume $P_i \subset C_i$ for any $i \in I$ if we are looking for the existence of non-trivial PR-perfect graphs..

3.2 Independent Assortment of Interests

We continue with our question of when a user-interest graph $Q(V, I, F)$ can be α-PR for α being an absolute constant. Informed by section 3.1, we consider directed user-interest graphs, where the production interests of a user are narrower than consumption interests. We ask: *What happens if users draw their interests at random and from the same distribution as all other users*? In other words, the interests are unstructured, so that user sets of different interests have little correlation.

Our results in this section are negative: it is not possible to achieve *constant* PR with high probability in such graphs, even with a separation of production and consumption interests. We show the result when every user has the same degree in Q, while interests could have non-uniform degrees. This result is fairly robust, and extends (under mild assumptions) to the case where users have non-uniform degrees (see Appendix A). This suggests that in a constant PR-perfect user-interest graph, every user does not draw her interests from the same distribution, and can be thought of as a necessary condition.

3.2.1 Random Regular Graphs

We now show that even with the separation of producers from consumers, it is not possible to achieve PR-perfection if the user-interest graph $Q(V, I, F)$ is generated at random. We begin with the observation that different interests have different cardinalities in terms of number of users. We plot the number of producers per interest for the 48 interests on Twitter in Fig. 3. We observe that some interests are much more popular than others.

We therefore consider a random graph model where the degree distribution W of I in $Q(V, I, F)$ need not be sharply

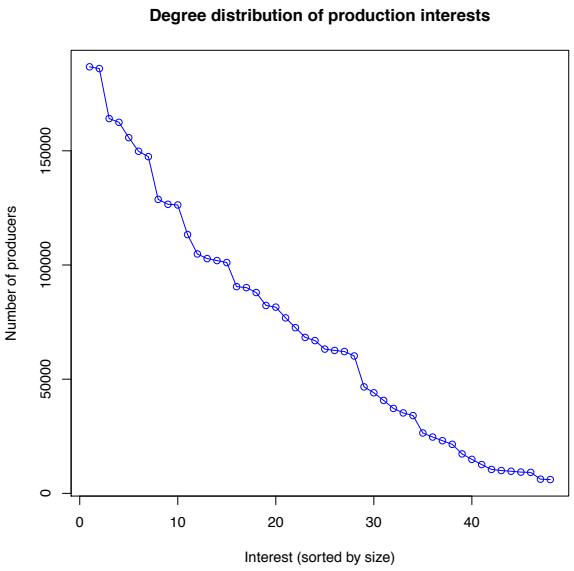

Figure 3: Number of Producers per Interest

concentrated. Our negative result holds under mild assumptions that the second moment of W is order of the mean. We use the following standard method of generating such a graph. Let $n = |V|$ denote the number of users, and $m = |I|$ denote the number of interests. We generate I as follows: We generate m_d interests with degree d, where $1 \leq d \leq n^\gamma$ for constant $\gamma < 1/3$, and $\sum_d dm_d = n$. We draw m_d from an appropriately scaled version of W, so that the mean is n and second moment is $O(n)$. A canonical example is Zipf(β) with $\beta > 3$, which follows the power law.

Let $C = |C(u)|$ and $P = |P(u)|$ denote the size of the produce and consume interests per user. Let I_d denote the interests in I with degree d. Consider a set I' where for every d, each interest $i \in I_d$ is replaced with d pseudo-interests of degree 1; note that $|I'| = \sum_d dm_d = n$. We generate the consume interests by choosing C random perfect matchings M between V and I' and taking the union of these matchings. Call this graph $M(V, I')$

After generating these matchings, consider any interest $i \in I$. Let C_i be the set of all users who were assigned to any of the d pseudo-interests corresponding to i. This defines the sets $C(u)$; note that $|C(u)| \leq C$ for all $u \in V$. Similarly, we generate the produce interests by choosing a random subset M' of size P of these matchings. Note that $|P(u)| \leq P$ for all $u \in V$. Call this graph of produce interests as $M'(V, I')$. Without loss of generality, we assume $|C(u)| = C$ and $|P(u)| = P$ – this is achieved by assigning the remaining interests arbitrarily. This defines the graph $Q(V, I, F)$.

THEOREM 3.3. *Suppose $P = \log^\delta n$ for constant $\delta > 2$, and $C \leq n^{1/12}$. Suppose further that the second moment of the degree distribution of I is upper bounded by the mean, and that the user-user graph $G(V, E)$ has maximum degree $o(n)$. Then, the graphs $Q(V, I, F)$ as generated above are not α-PR for any constant $\alpha > 0$, with high probability.*

PROOF. Since M and M' are regular random graphs, they are expanders with probability $1 - 1/n^2$. We will argue as

follows: Fix any user $u \in V$. For set $S \subseteq V$, let X_{uS} be an indicator random variable which is 1 if the following event happens: User u connects to set S **and** $|C(u) \cap (\cup_{v \in S} P(v))| \geq \alpha |C(u) \cup (\cup_{v \in S} P(v))|$. For realization r of the graph Q, let Y_{ur} be an indicator random variable which is 1 if user u can connect to some set $S \subseteq V$ such that $|C(u) \cap (\cup_{v \in S} P(v))| \geq \alpha |C(u) \cup (\cup_{v \in S} P(v))|$. Let Z_{urS} be an indicator random variable which is 1 if user u can connect to set $S \subseteq V$ such that $|C(u) \cap (\cup_{v \in S} P(v))| \geq \alpha |C(u) \cup (\cup_{v \in S} P(v))|$. Therefore, $\sum_{u,r} Y_{ur} \leq \sum_{u,r,S} Z_{urS} = \sum_{u,S} X_{uS}$. We will show that $\mathbf{E}[\sum_{u,S} X_{uS}] = o(1)$, which will imply $\mathbf{E}[\sum_{ur} Y_{ur}] = o(1)$. The latter quantity is an upper bound on the probability that a randomly chosen Q is α-PR, *i.e.*, whether there is some $G(V, E)$ so that every $u \in V$ is α-PR, which will complete the proof. Since the users are symmetric, we will simply fix a $u \in V$ and show that $\mathbf{E}[\sum_S X_S] = o(1/n)$ for this vertex.

Fix any $u \in V$. Note that $|C(u)| = C$. Consider $S = \{v_1, v_2, \ldots, v_r\}$, where $r = o(n)$. First note that since $M'(V, I')$ is an expander, with probability $1 - 1/n^2$, *every* set S of size at least $\sqrt{n}$ maps to at least $\sqrt{n}$ pseudo-interests, which must correspond to at least $\sqrt{n}/n^\gamma \geq n^{1/6}$ produce interests, since $\gamma < 1/3$. Therefore, for every such set, $|\cup_{v \in S} P(v)| = \omega(|C(u)|)$, so that, restricted to sets S of this size, $\sum_{r,S} Z_{rS} = \sum_S X_S = 0$ with probability at least $1 - 1/n^2$.

Therefore, we can restrict attention to sets S of size n^μ for $\mu < 1/2$. Fix some set $S = \{v_1, v_2, \ldots, v_r\}$. Each of these r nodes has $|P(v)| = P$, so that the total number of produce interests is at most $P \times r$. For interest $i \in I$, let L_i be an indicator random variable which is 1 if $i \notin C(u)$, but there exists v_j, $j \in \{1, 2, \ldots, r\}$ such that $i \in P(v_j)$. Therefore, $L_i = 1$ is a bad event corresponding to interest i contributing to the imprecision perceived by u. Recall that I_d is the subset of I with degree d. We have

$$\begin{aligned} \Pr[L_i = 1 | i \in I_d] &= \left(1 - \frac{dC}{n}\right)\left(1 - \left(1 - \frac{dP}{n}\right)^r\right) \\ &\approx \frac{rdP}{n}\left(1 - \frac{dC}{n}\right) \end{aligned}$$

The final approximation holds ignoring lower order terms as follows: $rdP = o(n)$ since $r, d = O(n^{1/2})$, and $P = \log^\delta n$. Therefore,

$$\begin{aligned} \mathbf{E}[\sum_i L_i] &\geq \sum_d \frac{dm_d}{n} rP - \sum_d \frac{d^2 m_d}{n} \frac{PCr}{n} \\ &= rP - o(1) \end{aligned}$$

To see the final equality, note that $\frac{\sum_d d^2 m_d}{n} = O(1)$, since this is the ratio of the second moment to the mean is $O(1)$ by assumption. Furthermore, $r < n^{1/2}$, $P = \log^\delta n$, and $C \leq n^{1/12}$. The variables L_i are negatively dependent, so that by an application of Chernoff bounds, for every small constant ϵ, we have:

$$\Pr\left[\sum_i L_i < rP(1 - \epsilon)\right] < e^{-rP\epsilon^2/2}$$

Therefore, the probability that there exists set $S = \{v_1, v_2, \ldots, v_r\}$ such that $\sum_i L_i < rP(1 - \epsilon)$ is at most $e^{r(\log n - P\epsilon^2/2)}$. This quantity is at most $\frac{1}{n^3}$ since $P = \Theta(\log^\delta n)$ for constant $\delta > 2$. If $\sum_i L_i < rP(1 - \epsilon)$, then u is not interested in a $(1 - \epsilon)$ fraction of the rP produce interests of set S, which implies u cannot be α-PR for constant α. This shows that $\mathbf{E}[\sum_S X_S] = o(1/n)$. Therefore, $Q(V, I, F)$ is not α-PR for any constant α with high probability. □

Though the above proof assumes each user has exactly Q consume interests and P produce interests, this assumption is not critical. The proof easily generalizes to distributions over degrees of users in $Q(V, I, F)$, as long as the degrees lie in $[\log^\delta n, n^\mu]$ for suitable constants $\delta > 2$ and $\mu < 1$.

3.2.2 Extension

The above result, though very strong, assumes users have super-constant number of interests and that this distribution is uniform. Our empirical analysis suggests that many users on Twitter have very few interests. In Appendix A, we consider a simple affiliation network model in the spirit of [7, 17] where users and interests have power law degree distributions (users for interests, and interests for users), and choose to associate independently subject to the degree constraints. We show that this model is *not* constant PR when the user-user graph densifies (or has super-constant average degree), an assumption that is widely believed to hold for social networks [18, 7]. It would therefore appear that the negative result is robust as long as users choose interests independently, so that there is little correlation between the user sets for different interests. It is important to note that our model is a simplification of affiliation networks, and the full model is quite powerful. So our result does not imply that affiliation networks are not a good model of social networks.

3.3 Empirical Analysis of a User-Interest Graph

The discussion so far has shown that to achieve constant PR, it cannot be that all users draw interests independently from the same distribution. We verify this condition on Twitter as follows. For each pair of the 48 interests described earlier (and listed in figure 4), we compute the number of users who are producers for both these interests. Let n_{ij} denote this value for interests i and j. Further, let n_i denote the total number of users producing interest i, and let $n = |V|$ denote the total number of users. Then, if the graph is formed by users repeatedly sampling from a common distribution, then in expectation, approximately $e_{ij} = n_j n_j / n$ users would produce both i and j. We compute the chi-squared measure:

$$\chi_{ij} = \frac{|n_{ij} - e_{ij}|}{\sqrt{e_{ij}}}$$

We next sort the χ values in decreasing order. Let W denote the graph on the $m = 48$ interests, where there is an edge between all pairs of nodes. Let W_p denote the graph that is obtained by only adding edges (i, j) between a fraction p of the node pairs with the largest χ_{ij} values. The graph W_p has the same density as the Erdos Renyi graph $G(m, p)$ does in expectation. We then compute the transitivity of W_p, which is the probability that two neighbors of a node are connected. For $p = 1/12$ and $p = 1/6$, these values are 0.63 and 0.61 respectively. This shows a very high degree of clustering compared to $G(m, p)$, whose average transitivity is approximately 0.09 and 0.17 for $p = 1/12$ and $p = 1/6$ respectively. The value $1/12$ is interesting because it is just larger than $\frac{\ln 48}{48}$ ($\frac{\ln n}{n}$ is the connectivity threshold for $G(m, p)$).

We then cluster the graph $W_{\frac{1}{12}}$ (using the fastgreedy method in R with default parameters), and show the clustering in

label	name	label	name
1	Music and Radio	2	Technology Industry
3	Politics	4	Sports
5	Photography	6	Adult
7	Technology	8	Baseball
9	Financial Services Industry	10	Travel Industry
11	Arts and Entertainment	12	Movie/Film/TV
13	International News	14	Sports
15	Football	16	Books
17	Healthcare Industry	18	Education
19	Retail Industry	20	Application Store
21	Fiction and Literature	22	Movie/Film/TV:Adult
23	Games	24	Fashion Industry
25	Professional Services Industry	26	Alcoholic Beverages
27	Specialty	28	Non-Profit
29	Racing	30	Online Sales
31	Advertising and Marketing	32	Soccer/Futbol
33	Specialty Store	34	Food
35	Magazine	36	Artists
37	DJs	38	Hip Hop/Rap
39	Software Developers	40	Business
41	Hockey	42	Consumer/Disposable Goods Industry
43	Mixed Martial Arts	44	Beauty & Personal Care
45	Real-Estate Industry	46	Boxing
47	Religion	48	Science

Figure 4: Topic labels for topics in Figure 5

Fig. 5. Note the emergence of several natural clusters, such as sports, technology, and journalism. We find the emergence of such a natural clustering of topics to be of independent interest, which needs further study. While some clusters are unsurprising (eg. sports), some others (eg. the cluster 3, 17, 18, 28, 35, 16, 21, 47) are non-obvious.

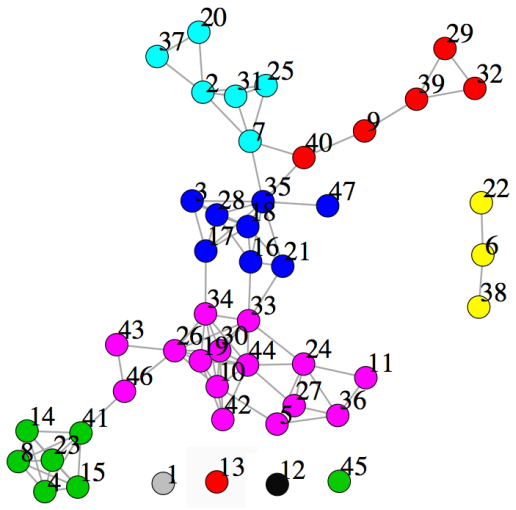

Figure 5: Communities of topics on Twitter. Yellow: Adult. Green: Sports. Pink: Consumer Retail. Dark Blue: Politics and News. Light Blue: Technology. Red: Financial.

4. ATTRIBUTE-BASED INTEREST MODEL

Our final theoretical result, somewhat surprisingly, is a positive one: we show that a natural and widely used generative model for interest-based social networks indeed achieves PR-perfection. This is the Kronecker graph model introduced in [18, 14, 21], where users are characterized by *attributes*, which are related to each other by a similarity measure. This model achieves several properties observed in social networks, such as power law degree distributions, shrinking diameter, and densification. We show a natural hierarchical definition of interests based on these attributes, which leads to PR-perfection and constant dissemination time.

4.1 Kronecker Graphs

In this model parametrized by a small number K, there are $|V| = n$ users, and $d = \log_K n$ attributes, each with K possible values from the set $S = \{a_1, a_2, \ldots, a_K\}$. Each node $u \in V$ maps to a d-dimensional vector of attribute values $(u_1, u_2, , u_d)$, where each $u_i \in S$. Therefore, $|V| = K^d = n$.

We define an interest as a set of pairs of attribute dimensions and their values, where a generic interest $i \in I$ has the following form:

$$i = \{\langle j_1, a_{j_1}\rangle, \langle j_2, a_{j_2}\rangle, \ldots, \langle j_r, a_{j_r}\rangle\}$$

$$\text{where } j_1, j_2, \ldots, j_r \leq K \text{ and } r \leq d$$

In other words, an interest is defined by specifying some $r \leq d$ attributes and their values. The set I is a subset of the set of all possible interests, so that $|I| \leq (K+1)^d$.

User-Interest Graph.

We now describe the mapping of users to producer and consumer interests. Treat the values in S as the K vertices of an undirected *seed graph* G_0, and denote the adjacency matrix of this graph as A. Assume $A[a_s, a_s] = 1$ for $1 \leq s \leq$

K. Consider interest $i = \{\langle j_1, a_{j_1}\rangle, \langle j_2, a_{j_2}\rangle, \ldots, \langle j_r, a_{j_r}\rangle\}$. The *consumers* of this interest are defined as:

$$C_i = \{u = (u_1, u_2, \ldots, u_d) \mid A[u_j, a_j] = 1 \ \ \forall \langle j, a_j\rangle \in i\}$$

In other words, for each component of i of the form $\langle j, a_j\rangle$, there must be an edge between u_j and a_j in G_0. Similarly, the producers of this interest are defined as:

$$P_i = \{u = (u_1, u_2, \ldots, u_d) \mid u_j = a_j \ \ \forall \langle j, a_j\rangle \in i\}$$

In other words, for each component of i of the form $\langle j, a_j\rangle$, the value u_j must coincide with a_j. It is clear that $P_i \subseteq C_i$ for all i.

The above interest model has the following interpretation. Since each interest is specified by a subset of attributes along with their values, the graph G_0 and adjacency matrix A specify which interests are related, i.e. which interests specify an *interested in* relationship. Further, the interests have a natural hierarchical structure, where the *broader* interests are those specified by fewer attributes. Also note that a producer of an interest needs to align with it's attribute values on all the relevant attribute dimensions, while a consumer of an interest only needs to be *interested in* those attribute values in the relevant attribute dimensions.

We can also derive further intuition about this interpretation by examining the typical size of these interest sets are. The size of interest sets depend on the nature of the adjacency matrix A. If the degree of each node a_i in G_0 is w and $|i| = d - j$, then $|P_i| = n^{\log w/\log K}(K/w)^j$. If we set $K = O(\log n)$, $w = O(1)$, and $j = O(1)$, then $|P_i| \approx n^{1/\log\log n}$. Furthermore, $|P_i|/|C_i| = 1/w^{d-j} = o(1)$, since $d = \log_K n \approx \frac{\log n}{\log\log n}$. This means the interest sets can be reasonably small (that is, $o(n^\gamma)$ for constant γ) for suitable choice of K, w, j, but within each interest set, we have a very small number of producers relative to consumers.

User-user Graph.

The graph $G(V, E)$ is undirected, and the generation process is the same as the one described in [21, 18]. For each $u = (u_1, u_2, \ldots, u_d)$ and $v = (v_1, v_2, \ldots, v_d)$, the edge (u, v) exists iff $A[u_j, v_j] = 1$ for all $j = 1, 2, \ldots, d$. In other words, two nodes connect iff they are interested in each other's attribute values on all attribute dimensions. It is shown in [18] that for suitably chosen adjacency matrices A, so that G_0 has constant diameter D, the graph $G(V, E)$ has multinomial degree distribution, has super-constant average degree (densifies) and the same constant diameter D as G_0. This therefore leads to a *densifying power law* graph G, and is termed the *Kronecker graph* on V using the attributes $\{1, 2, \ldots, d\}$ and seed graph G_0.

THEOREM 4.1. *Any user-interest graph $Q(V, I, F)$ and the associated user-user graph $G(V, E)$ generated by the above described process is PR-perfect with dissemination time at most $D + 1$.*

PROOF. Consider an arbitrary interest of the form $i = \{\langle j_1, a_{j_1}\rangle, \langle j_2, a_{j_2}\rangle, \ldots, \langle j_r, a_{j_r}\rangle\}$. Let $W = \{j_1, j_2, \ldots, j_r\}$, and $X = \{1, 2, \ldots, d\} \setminus W$. The set P_i has users u such that $u_j = a_j$ for $j \in W$. Consider the graph $G(P_i, E')$ induced on the set of users P_i. This graph is a Kronecker graph on the set P_i using the attributes X and seed graph G_0. This is therefore connected and has diameter at most D. This means any message originating at $u \in P_i$ reaches all of P_i in D hops. Further, it is easy to check that every neighbor of $u \in P_i$ is a $v \in C_i$, so that the precision is 100% and so is the recall. The total dissemination time is at most $D + 1$. □

Theorem 4.1 shows that there is indeed a model that achieves PR-perfection while preserving the key properties of social networks such as densification, heave tailed degree distribution, and shrinking diameter. The key aspects that made PR-perfection possible in Kronecker graphs are twofold: producers are a subset of consumers that are more aligned with that interest; and the interests have a hierarchical structure that enables users to connect to the appropriate producers. We have furthermore shown in previous sections that both these properties are necessary, including presenting empirical evidence validating their existence on Twitter.

4.2 Generalizing Kronecker Graphs

We now generalize the definition of interests in the Kronecker graph model to smoothly trade off the precision with the size of producer sets.

Consider the Kronecker graph model discussed in Section 4.1. We now show a broader definition of producers that leads to a smooth degradation in precision as the definition is broadened. Recall that K is the number of possible values an attribute can take, and that there are d attributes. Fix interest
$i = \{\langle j_1, a_{j_1}\rangle, \langle j_2, a_{j_2}\rangle, \ldots, \langle j_r, a_{j_r}\rangle\}$. The *consumers* of this interest are defined as:

$$C_i = \{u = (u_1, u_2, \ldots, u_d) \mid A[u_j, a_j] = 1 \ \ \forall \langle j, a_j\rangle \in i\}$$

We generalize the definition of a producer as follows. Consider user $u = (u_1, u_2, \ldots, u_d)$. For interest i, let $S_i = \{j_1, j_2, \ldots, j_s\}$ be a fixed set of at most s attributes. Then u produces i only if $u \in C_i$ *and* $\{j | u_j \neq a_j\}| \subseteq S$. This generalizes the definition of producers used in Section 4.1, which corresponds to setting $S = \phi$.

THEOREM 4.2. *For any $s \geq 0$, the Kronecker graph model with produce interests as defined above is α-PR for $\alpha = K^{-s}$. This is constant if K and s are constant.*

PROOF. From the discussion in Section 4.1, it is clear that any user u receives all interests in $C(u)$. Therefore, $|C(u) \cap S(u)| = |C(u)|$. In order to bound $|C(u) \cup S(u)|$, consider any interest $i \in C(u)$. Corresponding to i, there are at most K^s interests in $S(u)$ that are obtained by replacing the attributes in S_i with all possible values. There are K^s such interests. These interests could be produced by a neighbor of u, and hence be received by u though these need not belong to $C(u)$. Therefore, the graph is α-PR for $\alpha \geq K^{-s}$. □

A larger value of s implies producers are less aligned in attributes with the interest, *i.e.*, they are lower quality producers for that event. As is to be expected, this leads to a degradation in the precision of that interest. Therefore, this model shows a trade-off between the size of the producer set and the precision achieved, with smaller and more highly aligned producers leading to larger precision.

5. PRECISION ON TWITTER

Finally, we also present a preliminary empirical measurement of the precision observed on Twitter to build on the small user trial presented in Section 1. Since we cannot make

the direct measurement as was done in the user trial (we don't have the counter-factual of random tweets for users), we define the production and consumption interests using a procedure very similar to the one in Section 2.2. Namely, the set of consumption topics is obtained using the topic distribution for tweets that contain an URL that a user *clicked on*. And the set of production topics is obtained using the topic distribution for tweets that the user tweeted (or retweeted). Also as before, we sampled 1 million users from the 10 million highest PageRank users, and within these, restrict attention to those users *who have generated at least* $k = 10$ *tweets and clicks in a given 10 day period*. The rationale behind this was to avoid using dormant users and spammers in the analysis, and ensure that we have enough tweets to analyze the production and consumption distribution. We then select tweets that a user clicked on in their timeline in the given interval. The tags from the classifier give us the set of consumption topics for each user u, which is the same as $C(u)$. Similarly, we tag k uniform random tweets that a user v produced, giving a production distribution $P(v)$.

We indicate the set of edges between the users as E, and define an average empirical precision of user u as:

$$\text{Precision}(u) = \frac{\sum_{(u,v)\in E} |C(u) \cap P(v)|}{\sum_{(u,v)\in E} |P(v)|}$$

The formula above is an easy to compute approximation to an unbiased estimator constructed as follows: Each user u computes the multi-set $S(u) = \uplus\{P(v)|(u,v) \in E\}$. The precision seen by u is the probability that a randomly chosen interest in $S(u)$ belongs to $C(u)$. The reason for taking a multi-set union of the produce interests as opposed to a set union is that each user follows a large number (over a hundred) producers, and therefore, it is likely every interest is represented in one of these producers. Our estimator excludes *sparse* interests that are represented in only a few producers, from being counted towards precision.

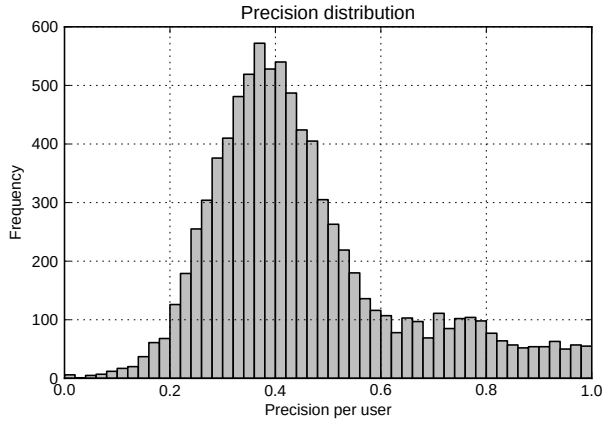

Figure 6: Distribution of Precision(u) on Twitter.

Using this measure, the average precision during the same time period was 40.5%, and the distribution of precision is presented in Fig. 6. As a baseline, since there are 48 interests and $\mathbf{E}[C(u)] \approx 8$, the average precision would be 17% had each consumer received all interests. The precision we obtain is significantly larger than the baseline, and we find this surprising given that we have only used a very narrow definition of consumption (clicks on tweets containing URLs). There are two possible explanations for this: The first is that tweets with URLs tend to be among the most interesting for users [1]; and the second is that we are measuring precision as the fraction of overlapping *interests* as opposed to the fraction of received tweets that are interesting – in this metric, we observe one in 2.5 received interests on any follow edge to be relevant on average. Nevertheless, it is clear that users read several tweets of interest without clicking on them, and as future work, we plan to determine better methods to measure precision empirically. We believe our user study provides a better and more reproducible template for performing such a study. We also re-emphasize that this measurement is not the central thesis of the paper, and is only provided as a preliminary datapoint of behavior on a real social network.

6. CONCLUSIONS AND FUTURE WORK

We have presented a definition of precision and recall for information dissemination on social networks using an interests-based framework. We also provide some necessary conditions on the structure of these interests to achieve good precision and recall, and validated these conditions on Twitter data. Somewhat surprisingly, we show that the Kronecker graph model achieves high precision and high recall while having constant dissemination time. We show preliminary empirical evidence towards the hypothesis that, despite widely held belief to the contrary, information flow on Twitter does indeed have high precision. Tying these together, the following explanation of this phenomenon emerges: users connect to other users based on similarity in interests, users produce content related to a narrower set of interests than they consume, and interests have structure so that users choose interests in a correlated fashion.

Our work is only a first step in understanding precision of information flow. Several research directions open up from this work. We have not really touched on recall or speed of dissemination, and it is *a priori* not even clear how to measure recall on Twitter. Furthermore, our measure of precision only uses a coarse set of interests, and the relation of the tweets to interests – in reality, even within an interest, tweets can have a wide range of "interestingness". This is harder to capture empirically, but is an interesting research direction. In a similar vein, we have not studied the structure of interests in Twitter in a very systematic way, since it is secondary to the main theme of this paper – this aspect will benefit from a more in-depth study.

Moving further afield, we have not considered the phenomena of *discoverability* and *coevolution*: Users need to discover other users who share their interests, and furthermore, users gradually change their links and the content they tweet based on the interests of their neighbors. These aspects need both theoretical modeling and empirical study.

7. ACKNOWLEDGMENTS

Ashish Goel acknowledges support from DARPA XDATA AND GRAPHS programs, and from the NSF award number 0904325. Kamesh Munagala is supported by an Alfred P. Sloan Research Fellowship, an award from Cisco, and by NSF grants CCF- 0745761, CCF-1008065, and IIS-0964560. Part of this work was done while

the author was visiting Twitter, Inc. The authors are also grateful to Alek Kolcz for his help with the classifier used in this work.

8. REFERENCES

[1] O. Alonso, C. Carson, D. Gerster, X. Ji, and S. U. Nabar. Detecting uninteresting content in text streams. In *SIGIR Crowdsourcing for Search Evaluation Workshop*, 2010.

[2] Apache. Apache Mahout, http://mahout.apache.org.

[3] D. Chakrabarti and C. Faloutsos. Graph mining: Laws, generators, and algorithms. *ACM Computing Surveys (CSUR)*, 38(1):2, 2006.

[4] P. Domingos and M. Richardson. Mining the network value of customers. *Proc. 7^{th} ACM SIGKDD Conference*, pages 57–66, 2001.

[5] D. Easley and J. M. Kleinberg. *Networks, crowds, and markets*, volume 8. Cambridge Univ Press, 2010.

[6] S. L .Feld. Why your friends have more friends than you do. *Amer. J. Sociology*, pages 1464–1477, 1991.

[7] I. Foudalis, K. Jain, C. H. Papadimitriou, and M. Sideri. Modeling social networks through user background and behavior. In *WAW*, pp 85–102, 2011.

[8] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proc. Nat. Acad. Sci.*, 99(12):7821–7826, 2002.

[9] M. S. Granovetter. The strength of weak ties. *Amer. J. Sociology*, pages 1360–1380, 1973.

[10] A. L. Hughes and L. Palen. Twitter adoption and use in mass convergence and emergency events. *Intl. J. Emergency Management*, 6(3):248–260, 2009.

[11] Twitter Inc. *Embedded Tweets*, 2013. `https://dev.twitter.com/docs/embedded-tweets`.

[12] R. Karp, C. Schindelhauer, S. Shenker, and B. Vocking. Randomized rumor spreading. In *Proc. 41st IEEE FOCS*, 2000.

[13] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. *Proc. 9th ACM SIGKDD Conf.*, 2003.

[14] M. Kim and J. Leskovec. Multiplicative attribute graph model of real-world networks. *Internet Mathematics*, 8(1-2):113–160, 2012.

[15] J. Kleinberg. Cascading Behavior in Networks: Algorithmic and Economic Issues. In N. Nisan, T. Roughgarden, E. Tardos, and V.V. Vazirani, editors, *Algorithmic Game Theory*. Cambridge University Press New York, NY, USA, 2007.

[16] J. M. Kleinberg. Navigation in a small world. *Nature*, 406(6798):845–845, 2000.

[17] S. Lattanzi and D. Sivakumar. Affiliation networks. In *Proc. 41st ACM STOC*, pages 427–434. 2009.

[18] J. Leskovec, D. Chakrabarti, J. Kleinberg, C. Faloutsos, and Z. Ghahramani. Kronecker graphs: An approach to modeling networks. *The J. Machine Learning Res.*, 11:985–1042, 2010.

[19] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. *ACM Trans. Web*, 1(1):5, 2007.

[20] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: Densification laws, shrinking diameters and possible explanations. In *Proc. 11th ACM SIGKDD Conference*, 2005.

[21] M. Mahdian and Y. Xu. Stochastic kronecker graphs. *Random Struct. Algorithms*, 38(4):453–466, 2011.

[22] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*, volume 1. Cambridge University Press, Cambridge, 2008.

[23] A. K. McCallum. Mallet: A machine learning for language toolkit. 2002.

[24] P. McFedries. Technically speaking: All a-twitter. *Spectrum, IEEE*, 44(10):84–84, 2007.

[25] S. Milgram. The small world problem. *Psychology Today*, 2(1):60–67, 1967.

[26] M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, 1(2):226–251, 2004.

[27] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999.

[28] N. Park, K. F. Kee, and S. Valenzuela. Being immersed in social networking environment: Facebook groups, uses and gratifications, and social outcomes. *CyberPsychology & Behavior*, 12(6):729–733, 2009.

[29] E. M. Rogers. *Diffusion of innovations*. Free Press, 2010.

[30] D. Sculley. Combined regression and ranking. In *Proc. 16th ACM SIGKDD Conference*, pages 979–988. 2010.

[31] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.

APPENDIX

A. SIMPLE AFFILIATION NETWORKS

The result in Section 3.2.1 assumes all users have similar number of interests, and this number is super-constant. However, on Twitter, we observe that the number of interests per user follows a skewed distribution plotted in Fig. 7.

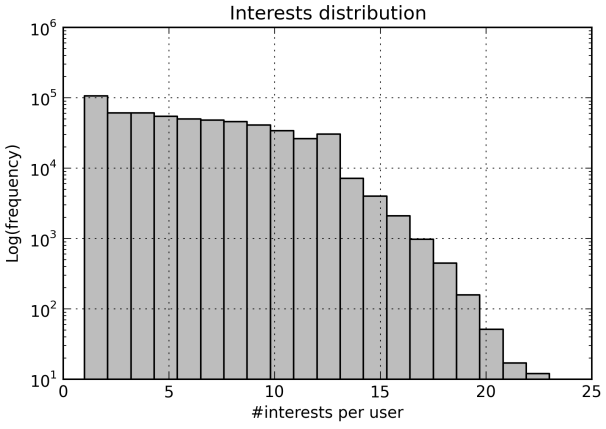

Figure 7: Histogram of number of interests per user

In order to model such behavior, we consider a fairly natural interest-based generative model of social networks termed affiliation network model [17, 7]. This model achieves many observed statistical properties of social networks (the graph $G(V, E)$), such as shrinking diameter, heave tailed de-

gree distributions, and super-constant average degree. Furthermore, it models both skewed interest degrees from Section 3.2.1, as well as the skew observed in Fig. 7.

The model we present simplifies the models presented in [17, 7], and we show this cannot be α-PR for any constant α, whenever the user-user graph $G(V, E)$ has super-constant average degree (or it densifies). Our model follows the discussion in Section 3.2.1 – the bipartite graph $Q(V, I, F)$ on users and interests is generated by the following random process. Fix two numbers $a_2 = 2 + \epsilon$, and $a_1 = 2 + 1/\epsilon$ for $0 < \epsilon < 1$. For instance, we can choose $a_2 = 2.5$ and $a_1 = 4$. There are n interest nodes I. The degrees in I are drawn from Zipf(a_2), with maximum degree n^γ for sufficiently small γ. The distribution X =Zipf(a) is integer valued, with $\Pr[X = r] \propto \frac{1}{r^a}$.

User-Interest Graph.

Now imagine there is an infinite pool of user nodes, whose degrees are drawn from Zipf(a_1). Again assume the maximum degree is n^γ for sufficiently small constant γ. For a user node u of degree $d(u)$, we split it into $d(u)$ *unit user* nodes each annotated with the degree $d(u)$. Each node $q \in Q$ of degree $d(q)$ chooses $d(q)$ unit user nodes uniformly at random from this infinite pool and connects to these nodes. The unit-user nodes that are connected to are considered *marked*. At the end of the process, we generate the final set of users V and the bipartite graph Q as follows: For every degree d, we collect together all marked unit-user nodes annotated with degree d. We group these nodes into buckets of size d, and each of these buckets becomes a user $u \in U$ with degree d. Note that there could be multiple parallel edges in $Q(V, I, F)$; we retain these edges for simplicity.

By a simple application of Chernoff bounds, the number of nodes in V with degree in $[1, n^\gamma]$ for sufficiently small γ agrees with the distribution Zipf(a_2) to within a factor of 2 w.h.p., and we ignore this error in the remaining discussion.

User-User Graph.

For any interest $i \in I$, let $V(i)$ denote the set of users having an edge to this interest. We generate the user-user graph $G(V, E)$ by *folding* the graph $Q(V, I, F)$ as follows: We place an edge between $u_1, u_2 \in V(i)$ with probability $1/r^{1-\delta}$, where $r = |V(i)|$ and $\delta > \epsilon$. The graph induced on $V(i)$ is therefore an Erdos-Renyi random graph $G(r, 1/r^{1-\delta})$.

It is shown in [17] that for the choice of parameters mentioned above, *i.e.*, $a_1 = 2+1/\epsilon$, $a_2 = 2+\epsilon$, and $0 < \epsilon < \delta \leq 1$, the resulting graph $G(V, E)$ has heavy-tailed degree distribution (since it stochastically dominates Zipf(a_1)), constant effective diameter for each interest set (since the induced graph is $G(r, 1/r^{1-\delta})$ for $\delta > 0$), and super-constant expected degree. This part requires $\delta > \epsilon$, and follows from an easy calculation that is implicit in the proof below. The canonical setting is to have $\delta = 1$, so that the graph induced on users sharing an interest is a complete graph.

We term the above model *Simple Affiliation Networks*. Since we are considering a generative process, we define the notion of expected precision:

DEFINITION A.1. *Given sets V, I, a generative process for $Q(V, I, F)$ and $G(V, E)$ is said to be α-EPR if*

$$\min_{v \in V} \frac{\mathbf{E}[|C(v) \cap S(v)|]}{\mathbf{E}[|C(v) \cup S(v)|]} \geq \alpha$$

where the expectation is over the process that generates Q, G.

We show the following theorem; our result holds for any $\delta > \epsilon$, so that the graph $G(V, E)$ densifies.

THEOREM A.2. *Assuming $|P_i| \geq 1$ for all interests i, the Simple Affiliation Network model with $\delta > \epsilon$ is not α-EPR for any constant α, regardless of the choice of P_i for each i.*

PROOF. We present the proof for the case $\delta = 1$. In the analysis below, we will focus on some user u, and condition on u having at degree (or number of interests) d. We will calculate the expected precision and recall of u conditioned on this event. Define the degree $d(\cdot)$ of a user (resp. interest) as their degrees in the bipartite graph $Q(V, I, F)$. Choose any $u \in U$, and let $|C(u)| = M \in [1, n^\gamma]$ for $\gamma < 1/10$. Now view the graph $Q(V, I, E)$ as follows: Each node $u \in V$ is $d(u)$ unit-user nodes of degree one, and each $i' \in I$ is $d(i')$ unit-nodes of degree 1. Therefore, we can view fixed $u \in V$ as M unit-nodes $u_1, u_2, \ldots, u_M$ each of which connects to an interest node at random. Fix some u_j. The interest node $i_j \in I$ connected to by u_j has degree $d(i_j)$ drawn from the distribution:

$$\Pr[\text{Degree of } i_j = d] = \frac{d \times 1/d^{a_2}}{\sum_{s=1}^{\infty} s \times 1/s^{a_2}}$$

Therefore, the degree distribution of the M users sharing interest i is Zipf($a_2 - 1$), where $a_1 = 1 + \epsilon < 2$. Consider some i_j with degree d_{i_j}, and consider some neighbor v of this interest. Then, with constant probability, the following two events happen for v: (1) Its degree is exactly 2; let the other interest shared by v be i'; and (2) the degree of i' is one, so that $v \in P(i')$. This means that for every neighbor v of i_j, with constant probability, v produces one interest. Node u receives this interest since $(u, v) \in E$ due to shared interest i_j. Therefore, the expected number of interests received by u due to interest i_j is $\Omega(d_{i_j})$. Since d_{i_j} is drawn from Zipf($a_2 - 1$), where $a_1 = 1 + \epsilon < 2$, the expected number of received interests of u is:

$$\mathbf{E}[d(u)] \left(\sum_{s=1}^{\infty} \Omega(s) \times \frac{1}{s^{a_2 - 1}} \right) \quad = \quad \omega(\mathbf{E}[d(u)])$$

Next note that $\mathbf{E}[d(u)] = \mathbf{E}[C(u)] = O(1)$ since the degree of u is distributed as Zipf(a_1). Therefore, $G(V, E)$ is not α-EPR for any constant α. □