Learning the nature of information in social networks

Rakesh Agrawal
Search Labs
Microsoft Research
rakesha@microsoft.com

Michalis Potamias*
Computer Science
Boston University
mp@cs.bu.edu

Evimaria Terzi*
Computer Science
Boston University
evimaria@cs.bu.edu

ABSTRACT

We postulate that the nature of information items plays a vital role in the observed spread of these items in a social network. We capture this intuition by proposing a model that assigns to every information item two parameters: *endogeneity* and *exogeneity*. The endogeneity of the item quantifies its tendency to spread primarily through the connections between nodes; the exogeneity quantifies its tendency to be acquired by the nodes, independently of the underlying network. We also extend this item-based model to take into account the openness of each node to new information. We quantify openness by introducing the *receptivity* of a node as an additional parameter in our model. Given a social network and data related to the ordering of adoption of information items by nodes, we develop a maximum-likelihood framework for estimating endogeneity, exogeneity and receptivity parameters. We apply our methodology to synthetic and real data and demonstrate the efficacy of our framework as a data-analytic tool.

1. INTRODUCTION

Since their introduction, online social networks have attracted millions of users, many of whom have integrated online social-network activity into their daily lives. This global phenomenon has inevitably increased the exposure of people to new information and affected the way information propagates. 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 [1, 2, 4, 6, 7, 16]. These studies rely on the assumption that it is the nature of the people, or the nature of the people's connections, that determines the form of information cascades.

While we recognize the impact of network structure and nodes' characteristics on information propagation we postulate 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. For example, consider the social network of people living in Boston. There, information about "global warming" is primarily exogenous, whereas information about "good physicians in the area" is primarily endogenous. Endogeneity and exogeneity need not be mutually exclusive; some items can be both highly endogenous as well as highly exogenous. For example, the propagation of the iPhone "Ocarina" application through friends has certainly played a key role in its huge popularity. However, many have discovered Ocarina by reading about it in various media.

In this paper, we define a probabilistic model that associates each information item i with two parameters: its *endogeneity* and its *exogeneity*. The endogeneity captures the effect of the network in the propagation of i. The exogeneity quantifies the extent to which the spread of i is due to factors external to the network. We further enhance our model with one more parameter per user to accommodate different users' behaviors; we call this parameter *receptivity*. Receptivity is a property of people, not of information items, and aims to capture each person's tendency to accept new information.

Given a social network and data related to the ordering of adoption of items by nodes, we develop a maximum-likelihood framework for learning the model parameters efficiently from the observed data. The obtained estimates do not depend on the granularity of the observation intervals, but rather on the sequence of the observed events. Using synthetic data we demonstrate that our learning methods accurately recover the parameters used for data generation. Finally, applying our framework on real data from online media sites we demonstrate its computational efficiency and its efficacy as a data-analysis tool.

Roadmap: After a brief overview of the related work in Section 2, we describe the item-based model and the corresponding parameter-estimation problem in Section 3. In Section 4 we describe our algorithm for solving this problem. We extend the basic item-based model with user receptivities in Section 5. We present experiments with real and synthetic datasets in Section 6 and conclude in Section 7.

2. RELATED WORK

We are not aware of any prior work that focuses on the characterization of information items as exogenous or endogenous based on their propagation patterns. Recent work on information propagation is focused on distinguishing between different types of peer effects. For example, Anagnostopoulos et al. [1] proposed a methodology for distinguishing between peer influence and peer correlation. Similarly,

*Work done at Search Labs during the summer of 2010.

Microsoft Research Technical Report.
MSR-TR-2011-59, May 6, 2011..

Aral et al. [2] considered the problem of separating influence-based contagion from homophily-driven diffusion. The conclusion of these studies is that the role of peer influence has been overestimated. However, in all these studies, all information items are considered to have identical propagation properties. We focus on unraveling the differences in the nature of the information items themselves.

A large body of work in algorithmic social-network analysis has focused on identifying *influential* nodes [9, 10, 11, 13]. Different definitions of influential nodes lead to different problems. But a common characteristic of all these approaches is that they *assume* a peer-influence model that defines the propagation of information in the network. Again, this propagation model is assumed to be identical for all items. Our work does not *assume* how information propagates in the network. Instead, we *learn* the specifics of information propagation from the observed data, allowing for each item to define its own propagation model.

Learning, rather than assuming, the strengths of influence between nodes was posed as a graph reconstruction problem by Mannila and Terzi [14]. Inferring the underlying influence network after observing the spread of a particular information item has also been studied by Rodriguez et al. [6]. Similarly, Goyal et al. [7] develop a framework for learning the influence probabilities between friends in a social network. All these methods quantify the influence between neighbors assuming that all items behave identically. Rather than focusing on identifying a global influence network, on which all items propagate, we focus on unraveling the specific propagation pattern of each item separately.

Rodriguez et al. [6] also include an *external source* to their model (similar to our exogenous node) to explain cascade jumps; they set the probability of the edges of that node to a small ϵ. Instead, we use one such node per item and learn its influence probability; besides, our experiments indicate that the distribution of the exogeneity is skewed amongst items.

Analysis of social-media data with respect to the information items has appeared in the work of Gruhl et al. [8] and, more recently, in the work of Mathioudakis et al. [15]. The focus of Gruhl et al. [8] is on topic modeling, and their goal is to identify long-running and spiky topics. The problem of grouping information items into topics is orthogonal to our work. Mathioudakis et al. [15] work in an online setting and their goal is to predict which items will attract large attention. Although we mainly focus on offline characterization of items, we also perform prediction experiments to validate our model.

3. THE ITEM-BASED MODEL E2

In this section, we present our basic model and the corresponding parameter-estimation problems. First, we present some preliminaries and some notational conventions.

3.1 Preliminaries

We assume that the social network is a directed graph $G = (V, E)$ with $|V| = n$ users. There is a link between two nodes $u, u' \in V$, denoted by $(u \to u')$ if node u *follows* node u'. Such a directed link suggests that there is potential influence (i.e., propagation of information) from u' to u. Throughout the paper, we consider all links as equally significant[1]. We also 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. Every element of the vector is associated with an information item $i \in \{1 \dots m\}$. We write $A_u^t(i) = 1$ 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. We also assume that once a node becomes active with respect to an item, then it remains active; this assumption is also known as *progressive item propagation* [10]. Finally, we use $\mathbf{A}$ to denote the $n \times m$ matrix that encodes the *observed* activation state at the end of the observation period. In other words, $\mathbf{A}(u, i) = 1$ if node u has, at some point, become active with respect to item i. Otherwise, $\mathbf{A}(u, i) = 0$. We call this matrix the *activation matrix*.

In addition to the activation matrix $\mathbf{A}$, we also use the $n \times m$ *active-neighborhood* matrix $\mathbf{\Gamma}$. Element $\mathbf{\Gamma}(u, i)$ 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. Given graph $G = (V, E)$ and the sequence of activations encoded in vectors A_u^t, we can construct the active-neighborhood matrix $\mathbf{\Gamma}$ in $\mathcal{O}(nm)$ time.

3.2 The E2 model

We claim that the propagation of an information item in a given network depends on the item's nature. More specifically, every item $i \in \mathcal{I}$ is characterized by two parameters: its *endogeneity*, $e_i \in [0, 1]$ and its *exogeneity*, $x_i \in [0, 1]$. That is, every item i is characterized by the pair of parameters $\theta_i = (e_i, x_i)$. The endogeneity parameter characterizes the item's tendency to propagate through the network due to the peer effect. The exogeneity parameter captures the item's tendency to be independently generated by nodes in the network. This model is consistent with our thesis that not all items are equally endogenous and equally exogenous.

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. Note that our model assumes that each active neighbor tries to independently activate u. At the end of the observation period, u becomes active with respect to i, with probability: $1 - (1 - x_i)(1 - e_i)^{\mathbf{\Gamma}(u,i)}$.

We call this item-based model the E2 model. For a dataset with m items, there is a total of $2m$ parameters. We use $\mathbf{e}$ and $\mathbf{x}$ to represent the vectors of all items' endogeneity and exogeneity parameters. We use $\mathbf{\Theta} = \langle \mathbf{e}, \mathbf{x} \rangle$ to denote the vector of these pairs of values for all items.

Generative model: 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

[1] Our methods can be generalized to the case where these links are associated with weights (e.g., influence probabilities). Such weights can be obtained via existing techniques [6, 7].

directed links $(u \to s_i)$. 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 process proceeds in discrete steps. At each time step, activation of any node u, through links $(u \to s_i)$, 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. Thus, there can be at most $|E_i|$ activation attempts. The final activation state of all nodes with respect to all items is stored in the final activation matrix $\mathbf{A}$.

The assumption that every active node has a single chance to propagate an item to each one of its inactive neighbors is common in many propagation models e.g., Independent Cascade (IC) model [5, 10]. In IC, every active node is given a single chance to activate any of its inactive neighbors. The success of the activation is a model parameter and depends, in principle, on the pair of nodes and the information item. IC can simulate E2 if executed on the activation graph H_i with parameters θ_i.

3.3 Problem Definition

Given the active-neighborhood information $\mathbf{\Gamma}$ and parameters $\mathbf{\Theta}$, we can compute the likelihood of the observed activation matrix $\mathbf{A}$ as follows:

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

The probability that node u becomes active with respect to i given that u has $\mathbf{\Gamma}(u,i)$ active friends can be computed using the basic assumptions of the E2 model as follows:

$$\begin{aligned}\Pr\left(\mathbf{A}(u,i) = 1 \mid \mathbf{\Gamma}(u,i), e_i, x_i\right) = \\ = \quad 1 - (1-x_i)(1-e_i)^{\mathbf{\Gamma}(u,i)}.\end{aligned}$$

Naturally,

$$\begin{aligned}\Pr\left(\mathbf{A}(u,i) = 0 \mid \mathbf{\Gamma}(u,i), e_i, x_i\right) = \\ 1 - \Pr\left(\mathbf{A}(u,i) = 1 \mid \mathbf{\Gamma}(u,i), e_i, x_i\right).\end{aligned}$$

Our goal is to estimate parameters $\theta_i = (e_i, x_i)$ for every item. In other words, 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:

$$\mathcal{L}\left(\mathbf{A} \mid \mathbf{\Gamma}, \mathbf{\Theta}\right) \quad = \quad \log \Pr\left(\mathbf{A} \mid \mathbf{\Gamma}, \mathbf{\Theta}\right). \quad (2)$$

We call this problem the ML-E2 problem and we formally define it as follows:

PROBLEM 1 (ML-E2). *Given activation matrix* $\mathbf{A}$ *and the active-neighborhood matrix* $\mathbf{\Gamma}$*, find parameters* $\mathbf{\Theta} = \langle \mathbf{e}, \mathbf{x} \rangle$*, such that the loglikelihood of the observed activations, given these parameters, is maximized, i.e.,*

$$\mathbf{\Theta} = \arg\max_{\mathbf{\Theta}'} \mathcal{L}\left(\mathbf{A} \mid \mathbf{\Gamma}, \mathbf{\Theta}'\right). \quad (3)$$

4. PARAMETER ESTIMATION

In this section, we present an efficient method to solve Problem 1. We also demonstrate how analytic manipulations of the loglikelihood function (Equation (2)) can lead to significant speedups.

Using Equation (1), we rewrite the likelihood as

$$\mathcal{L}\left(\mathbf{A} \mid \mathbf{\Gamma}, \mathbf{\Theta}\right) = \sum_{i \in \mathcal{I}} \sum_{u \in V} \log\left(\Pr\left(\mathbf{A}(u,i) \mid \mathbf{\Gamma}(u,i), e_i, x_i\right)\right). \quad (4)$$

If we use $L_i(e_i, x_i)$ to represent the quantity

$$L_i(e_i, x_i) \triangleq \sum_{u \in V} \log\left(\Pr\left(\mathbf{A}(u,i) \mid \mathbf{\Gamma}(u,i), e_i, x_i\right)\right), \quad (5)$$

then we can rewrite the likelihood as:

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

This implies that 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. Therefore, an off-the-shelf optimization method (e.g., Newton Raphson method [17]) can be used to efficiently find the optimal values of the parameters. In our experiments we initialized our optimization routine from multiple starting points. However, we always computed the same parameters regardless of the starting point.

We refer to the method of estimating the $2m$ items' parameters as the `Item-Fit` algorithm. The running time of `Item-Fit` is $\mathcal{O}(mT)$, where T is the time required by the optimization algorithm to find the optimal values of (e_i, x_i) of a single item i. Common optimization algorithms are iterative. In every iteration t there is an estimate of the parameters $(e_i^{(t)}, x_i^{(t)})$. Given these estimates the algorithm recomputes the value of $L_i(e_i^{(t)}, x_i^{(t)})$ as well as the value of its gradient and uses these values to update parameters $(e_i^{(t+1)}, x_i^{(t+1)})$. Computing the value of $L_i(e_i^{(t)}, x_i^{(t)})$ and its gradient using the evaluation procedure implied by Equation (5) takes $\mathcal{O}(n)$ time per iteration.

Speeding up `Item-Fit`: The sparsity of real data allows us to speedup the computation of $L_i(e_i, x_i)$ and its gradient. Observe that $\Pr\left(\mathbf{A}(u,i) = 1 \mid \mathbf{\Gamma}(u,i), e_i, x_i\right)$ is the same for all active (resp. inactive) users that have the same number of active neighbors at the time of their activation (resp. end of observation period). For item i, consider all active nodes that had $\mathbf{\Gamma}(u,i) = \gamma$ active neighbors when they themselves became active. If we use $B^i(\gamma)$ to denote this set of active nodes, then for every $u \in B^i(\gamma)$

$$\begin{aligned}P_\gamma \quad &\triangleq \quad \Pr\left(\mathbf{A}(u,i) = 1 \mid \mathbf{\Gamma}(u,i) = \gamma, e_i, x_i\right) \\ &= \quad 1 - (1-x_i)(1-e_i)^\gamma.\end{aligned}$$

For item i, we also use $\hat{B}^i(\gamma)$ to refer to the set of inactive users that have γ active neighbors. Then, for each $u \in \hat{B}^i(\gamma)$

$$\Pr\left(\mathbf{A}(u,i) = 0 \mid \mathbf{\Gamma}(u,i) = \gamma, e_i, x_i\right) = 1 - P_\gamma.$$

We rewrite Equation (5) as follows:

$$\begin{aligned}L_i(e_i, x_i) = \quad & (6) \\ \sum_\gamma \Big(\sum_{u \in B^i(\gamma)} \log\left(P_\gamma\right) + \sum_{u \in \hat{B}^i(\gamma)} \log\left(1 - P_\gamma\right) \Big) \\ = \quad \sum_\gamma \Big(\left|B^i(\gamma)\right| \log\left(P_\gamma\right) + \left|\hat{B}^i(\gamma)\right| \log\left(1 - P_\gamma\right) \Big).\end{aligned}$$

Using Equation (6) for evaluating the value of $L_i(e_i, x_i)$ within every iteration of the optimization procedure results

to processing time that depends on the distinct values of γ that appear in the dataset. The number of distinct values of γ that occur for each item is a small constant. So using this transformation we achieve time $\mathcal{O}(1)$ per iteration. We remark that the optimization routine rarely exceeded 10 iterations in our experiments. Therefore, the overall running time is, in practice, linear to the number of items.

5. THE E2R MODEL

Although E2 captures the observed variation between the items, it does not capture that different nodes may react differently to the same information item. In this section, we present the E2R model which incorporates one *receptivity* parameter per user.

Every node u is associated with a single parameter $r_u \in [0,1]$; this parameter 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. We use $\mathbf{r}$ to denote the n-dimensional vector with the receptivity parameters of all users.

We incorporate the user-receptivity parameters in the E2 model. The probability of the observed activation matrix $\mathbf{A}$ given the item parameters $\mathbf{\Theta}$ and user receptivities $\mathbf{r}$ is given by:

$$\Pr(\mathbf{A} \mid \mathbf{\Gamma}, \mathbf{\Theta}, \mathbf{r}) = \prod_{i \in \mathcal{I}, u \in V} \Pr(\mathbf{A}(u,i) \mid \mathbf{\Gamma}(u,i), e_i, x_i, r_u). \quad (7)$$

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

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

The sole difference between $\Pr(\mathbf{A} \mid \mathbf{\Gamma}, \mathbf{\Theta}, \mathbf{r})$ and $\Pr(\mathbf{A} \mid \mathbf{\Gamma}, \mathbf{\Theta})$ is that for every user u and item i, terms x_i and e_i are multiplied by the factor r_u. Intuitively, every time we have an endogenous or exogenous attempt to activate a user, the user also needs to accept that activation. Clearly, the E2R model is equivalent to the E2 model when $r_u = 1$ for every node u.

Using Equation (7), the *loglikelihood* of E2R is

$$\begin{aligned} \mathcal{L}_R(\mathbf{A} \mid \mathbf{\Gamma}, \mathbf{\Theta}, \mathbf{r}) &= \log(\Pr(\mathbf{A} \mid \mathbf{\Gamma}, \mathbf{\Theta}, \mathbf{r})) \\ &= \sum_{i \in \mathcal{I}, u \in V} \log(\Pr(\mathbf{A}(u,i) \mid \mathbf{\Gamma}(u,i), e_i, x_i, r_u)). \end{aligned} \quad (8)$$

Modeling power of receptivity: 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.

We now define the maximum-likelihood estimation problem for the E2R model.

PROBLEM 2 (ML-E2R). *Given activation matrix* $\mathbf{A}$ *and active-neighborhood matrix* $\mathbf{\Gamma}$*, find parameters* $\mathbf{\Theta} = \langle \mathbf{e}, \mathbf{x} \rangle$ *and* $\mathbf{r}$ *such that the loglikelihood of the data given these parameters is maximized, i.e.,*

$$\langle \mathbf{\Theta}, \mathbf{r} \rangle = \arg\max_{\mathbf{\Theta}', \mathbf{r}'} \mathcal{L}_R(\mathbf{A} \mid \mathbf{\Gamma}, \mathbf{\Theta}', \mathbf{r}'). \quad (9)$$

Disentangling the effect of dependent factors in information propagation is challenging. Accordingly, the ML-E2R is harder than ML-E2. In fact, the addition of users' receptivity yields a high dimensional optimization function ($2m + n$ parameters); this problem is hard to solve optimally. Therefore, we describe an efficient coordinate descent procedure to find a local optimum for Problem 2. Our solution builds upon the parameter-estimation methodology developed in Section 4.

The core of our method for solving this problem, which we call `ItemUser-Fit`, is the following: if parameters $\mathbf{r}$ are fixed we have a two-variable optimization problem similar to ML-E2 and we can use `Item-Fit` presented in Section 4 to efficiently estimate parameters $\mathbf{\Theta}$. Similarly, if parameters $\mathbf{\Theta}$ are fixed, we can efficiently estimate the user parameters $\mathbf{r}$. This last step can be done by optimizing $\mathcal{L}_R$ where the only variables are users' receptivities. Since users are independent, this estimation can be done separately, this time for every user, by solving a single-variable optimization problem. Algorithm `User-Fit` iterates between the two convex problems. Since the value of the likelihood decreases at each iteration, `ItemUser-Fit` converges to a local optimum. However, the speedup of `Item-Fit` cannot be used due to the mixing of the users' and the items' parameters. In practice, `ItemUser-Fit` handled our dataset efficiently and it converged in a dozen iterations.

6. EXPERIMENTS

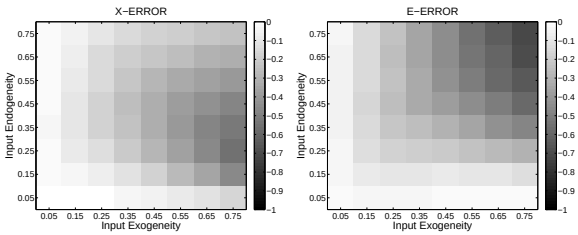
In this section, we present experimental results using real and synthetic datasets. First, we show that our algorithms accurately recover the model parameters using synthetic data. Next, we report interesting quantitative and qualitative results and demonstrate the fit of our model to real data. We implemented our solution using C++ and Matlab's optimization toolbox. We ran our experiments on 64-bit Intel Xeon 2.66 GHz, 6 GB RAM running Windows 7.

6.1 Experiments with synthetic data

Our experiments with synthetic data illustrate that `Item-Fit` recovers exogeneity and endogeneity values that are very close to the values used for data generation.

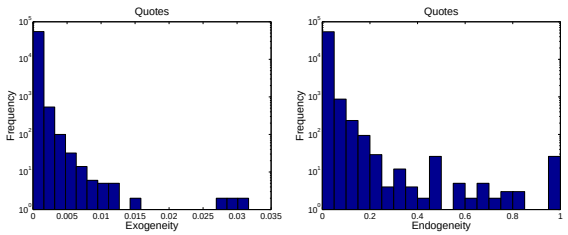
Data-generation. Since many real world datasets, including the data we analyze, have power-law degree distribution, we generate synthetic `ScaleFree` graphs using the model proposed by Barabási and Albert [3]. For every item i we pick uniformly at random an endogeneity parameter e_i in the interval $[0, 0.8]$ and an exogeneity parameter x_i from the interval $[0, 0.8]$. The graphs generated by this model are undirected. We convert them into directed graphs by creating two directed edges for every undirected edge.

Next, we run the following propagation procedure: For every item i we create the activation graph $H_i = (V \cup \{s_i\}, E_i)$ as described in Section 3 (recall that s_i is the exogenous node). During the propagation, we maintain a set R of edges that are ready to cause activation. All the edges $(v \to u)$ in R consist of *active node* u and *inactive node* v. Initially,



(a) X-Error (b) E-Error

Figure 1: `ScaleFree` graphs with $n = 1000$ number of nodes, 1% density, and $m = 1000$ items. Figure 1(a): X-Error for various combinations of input parameters $\mathbf{x}$ and $\mathbf{e}$. Figure 1(b): E-Error for various combinations of input parameters $\mathbf{x}$ and $\mathbf{e}$. Lighter cell color corresponds to smaller error.



(a) Exogeneity (b) Endogeneity

Figure 2: E2 model on `MemeTracker` dataset; y-axis in *logarithmic* scale. Figure 2(a): histogram of exogeneity of quotes. Figure 2(b): histogram of endogeneity of quotes (the scale on the x-axis is different in the two histograms).

only node s_i is active and R contains $|V|$ edges $(u \to s_i)$ for all $u \in V$. At every step, the data-generation process picks uniformly at random, without replacement, an edge $(u \to v)$ from R. If the active node v is node s_i then the activation of u succeeds with probability x_i. If the active node v is an active node from V, then the activation of u succeeds with probability e_i. If the activation of u succeeds, we set $\mathbf{A}(u,i) = 1$ and $\mathbf{\Gamma}(u,i)$ is set to be the number of active neighbors of u that are in V. Also, all edges $(w \to u)$ such that w is inactive are inserted in set R. The data-generation process stops when R becomes empty.

Evaluation metrics. Denote $\widehat{\theta}_i = (\widehat{e}_i, \widehat{x}_i)$ the recovered parameters for every item i. For the set of m observed items $\mathcal{I}$, we quantify the quality of the recovery using the average absolute recovery error. We define the *exogeneity absolute error* for the exogeneity parameters as

$$\text{X-Error}(\mathbf{\Theta}, \widehat{\mathbf{\Theta}}) \quad = \quad \frac{1}{m} \sum_{i \in \mathcal{I}} |x_i - \widehat{x}_i| \,.$$

The *endogeneity absolute error*, E-Error, is defined similarly.

Results. The `ScaleFree` graphs of this experiment have $n = 1000$ nodes, $m = 1000$ items, and density equal to 1% – each node has at least 5 neighbors. The results for graph-densities in the range $[0.4, 4]$% are similar. Also, varying the number of nodes and the number of items does not affect the recovery results. For each set of exogeneity/endogeneity parameters we generate 30 independent graphs and average the results.

Figure 1 shows the X-Error and E-Error values with respect to the range of values from which the data-generation parameters $\mathbf{x}$ and $\mathbf{e}$ are selected. Every interval in the x-axis (y-axis) corresponds to the interval from which the exogeneity (endogeneity) parameter was sampled uniformly at random. The average value of each interval is shown on the axes. Darker colors correspond to larger average errors.

Observe 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. Even if *all* these activations are due to exogeneity the exogeneity value cannot be larger than 0.1. Thus, the information items always fall in the leftmost column of the plots presented in Figure 1. There, both X-Error and E-Error are negligible.

6.2 Experiments with `MemeTracker` data

We next turn our attention to the practical utility of our models in real data. Apart from giving several insightful quantitative and anecdotal findings that highlight the efficacy of our model as a data analysis tool, we also validate the fit of our data via a prediction experiment after appropriately splitting the data into training and test sets.

Description of the data. We use the memetracker data [12]. This dataset 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. Finally, there are directed hyperlinks among posts.

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 $(b \to b')$ 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 denote this set by $\mathcal{Q}$. We say that blog u became active with respect to quote $q \in \mathcal{Q}$ at time t, if t was the first timestamp that u used q in one of his blogposts. We refer to the dataset consisting of graph $G_B = (V_B, E_B)$ along with the set of quotes $\mathcal{Q}$ as the `MemeTracker` dataset. It consists of 49373 distinct blogs and 56888 quotes. There are 171653 directed links among blogs and 605777 activations of blogs with respect to quotes (i.e., appearances of the quotes on the blogs). The distributions of the number of blogs each quote appears in, the number of quotes per blog, and the number of incoming and outgoing links per blog, resemble power-laws.

6.2.1 *Analyzing* `MemeTracker` *data using* E2

We use the `Item-Fit` algorithm to compute the endogeneity and exogeneity parameters for all the quotes in `MemeTracker`. Running `Item-Fit` on `MemeTracker` takes about 3 minutes.

Distribution of Endogeneity and Exogeneity Values. In Figure 2, we present the distribution of endogeneity and exogeneity values of the quotes. Note that in both plots the y-axis is in log-scale, and also that the x-axes are scaled differently. The skewed distribution of both exogeneity and endogeneity values shows that a non-negligible number of items are much more endogenous/exogenous than most items.

Most of the quotes have a small exogeneity value, while the maximum value of exogeneity on the x-axis is 0.032.

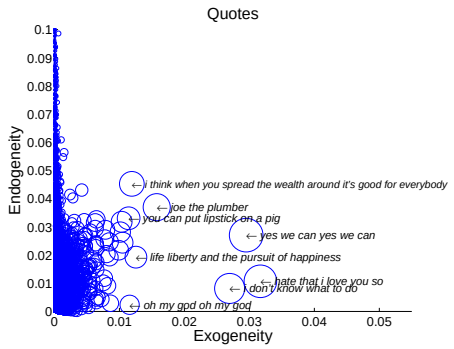


Figure 3: E2 model on `MemeTracker` dataset: x-axis: exogeneity of quotes. y-axis: endogeneity of quotes and marker area proportional to frequency.

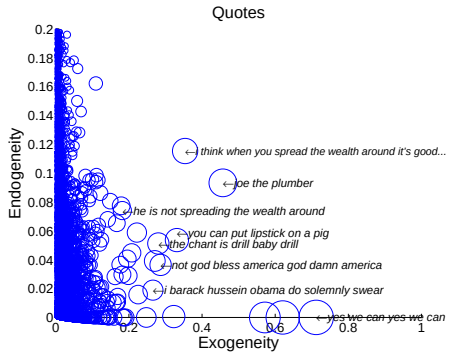


Figure 4: E2R model on `MemeTracker` dataset: x-axis: exogeneity of quotes, y-axis: endogeneity of quotes and marker area proportional to frequency.

Figure 2(b) shows a histogram of the endogeneity values. Contrary to the exogeneity values, the endogeneity in Figure 2(b) spans the whole range $[0, 1]$. Some quotes of high endogeneity propagate through most network edges they come across. These are infrequent quotes that propagate in small, isolated parts of the network. Therefore, they are highly endogenous.

Figure 3 is a scatter-plot of the exogeneity (x-axis) and the endogeneity (y-axis) of the quotes. The area covered by each marker is proportional to the frequency of the quote (i.e., the number of nodes it appears). We have also printed frequent dominant quotes. Observe that there are many items that exhibit high endogeneity and small exogeneity. Exogeneity appears to correlate well with frequency. This is expected since exogeneity is the probability that a random isolated node has a particular item; such a probability can only be large for frequent items.

Anecdotal results. In Figure 3, contrast the popular quote of Barack Obama in the 2008 U.S. presidential elections `i think when you spread the wealth around it's good for everybody` with the phrase from the U.S. Declaration of Independence `life liberty and the pursuit of happiness`. They both have similar exogeneity but the former is far more endogenous than the latter. We conjecture that this is due to the fact that the former was a *new* quote – as such it heavily used the network to propagate. The latter is a classic phrase, often used without triggering any cascade.

Table 1: `MemeTracker` dataset. Top-5 frequent quotes with **Exogeneity=H** that correspond to buckets (H,H) and (H,L) of the equi-depth (X,E)-histogram.

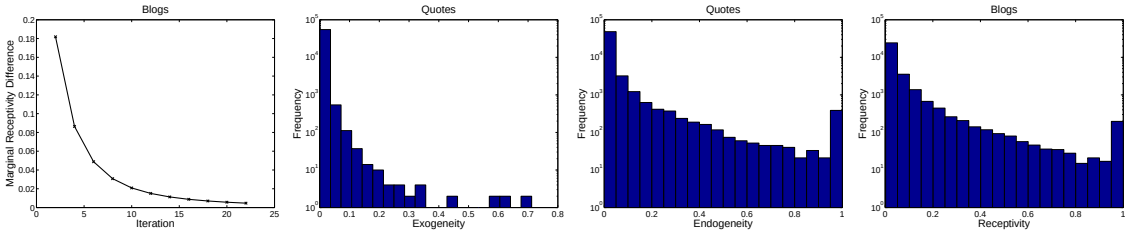
	Endogeneity=H **Top-5 frequent quotes from (H,H) bucket**
1.	yes we can yes we can
2.	hate that i love you so
3.	joe the plumber
4.	i think when you spread the wealth around it's good for everybody
5.	you can put lipstick on a pig
	Endogeneity=L **Top-5 frequent quotes from (H,L) bucket**
1.	i don't 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

Table 2: `MemeTracker` dataset. Top-5 frequent quotes with **Exogeneity=L** that correspond to buckets (L,H) and (L,L) of the equi-depth (X,E)-histogram.

	Endogeneity=H **Top-5 frequent quotes from (L,H) bucket**
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
	Endogeneity=L **Top-5 frequent quotes from (L,L) bucket**
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 neighbor
4.	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...

In order to present anecdotal findings in a principled manner we use the estimated endogeneity and exogeneity values of every quote to construct a 2-dimensional, 2×2 equi-depth histogram of the quotes. We call this 2-dimensional histogram the (X, E)-histogram. Using "H" ("L") to represent "High" ("Low") values of exogeneity or endogeneity, the 4 buckets of the equi-depth histogram correspond to the combinations (H,H), (H,L), (L,H) and (L,L) of values that (exogeneity, endogeneity) pairs take. Every bucket described by such a pair contains quotes with the same cumulative frequency. Within every bucket, we sort the quotes in decreasing frequency.

Table 1 shows the top-5 frequent quotes from buckets with high exogeneity, that is, buckets (H,H) and (H,L). Table 2 shows the top-5 frequent quotes from buckets with low exogeneity, that is, buckets (L,H) and (L,L).



(a) Convergence (b) Exogeneity (c) Endogeneity (d) Receptivity

Figure 5: E2R model on `MemeTracker` dataset; Figure 5(a): Convergence of the `ItemUser-Fit` algorithm. x-axis: iteration number of `ItemUser-Fit`; y-axis: marginal difference in the average receptivity estimates between rounds. Figures 5(b) 5(c) 5(d): y-axis: frequency in *logarithmic* scale. Figure 5(b): histogram of exogeneity of quotes. Figure 5(c): histogram of endogeneity of quotes. Figure 5(d): histogram of receptivity of blogs.

Before we go into more detail we make two observations: first, quotes with "Exogeneity=H" (Table 1) exhibit shorter length than quotes with "Exogeneity=L" (Table 2). Second, a web search reveals that most quotes with "Endogeneity=H" (upper part of both tables) were news-stories or popular quotes of the observation period.

Amongst the high-exogeneity quotes shown in Table 1 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. Contrast these quotes with the (H,L) quotes reported in the lower part of Table 1. 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 shown in Table 2 we can again distinguish between those for which "Endogeneity=H" and those with "Endogeneity=L". The first (shown in the upper part of Table 2) 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), shown in the lower part of Table 2. 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.

6.2.2 *Analyzing* `MemeTracker` *data using* E2R

After exploring the utility of E2 as a data analysis tool we move on to explore the E2R model.

The parameter estimation for the E2R model was made using the `ItemUser-Fit` algorithm. It took about two hours for the algorithm to converge on a single-thread implementation. In our experiment with the `MemeTracker` dataset the algorithm converged after a small number of iterations, which shows that the algorithm is quite usable in practice despite its iterative nature. Figure 5(a) illustrates the marginal average difference in the receptivity values obtained per iteration. We plot the first 22 iterations. Observe that changes are negligible after a dozen iterations. The convergence plots for the endogeneity and exogeneity parameters are similar and not shown.

The distributions of exogeneity and endogeneity using the E2R model are shown in Figures 5(b) and 5(c). The histogram of receptivity of nodes is shown in Figure 5(d). In all three plots the y-axis is in logarithmic scale. Observe that the distribution of exogeneity in E2R is almost identical to the distribution of exogeneity in E2 (see Figure 2(a)). Their correlation coefficient is 0.99 indicating that the introduction of receptivity hardly affected the exogeneity. On the other hand, the correlation between the E2 and E2R endogeneity values is equal to 0.66. This indicates that the introduction of receptivity has an effect on the endogeneity.

Anecdotal results obtained from the E2R model are shown in Figure 4. This is a scatter-plot similar to the one depicted in Figure 3. There are some changes in the anecdotes of the (X,E)-histogram: Some frequent quotes, namely, `yes we can`, `life liberty and the pursuit of happiness`, and `hate that I love you so` moved from bucket (H,H) to the low endogeneity bucket (H,L). Interestingly, in contrast with E2, E2R found negligible endogeneity for these popular quotes (note that all lie on the x-axis of Figure 4). These quotes did *not* propagate to nodes that were highly receptive with other quotes. The inference procedure of E2R penalizes their endogeneity to capture that fact. Indeed, `yes we can, yes we can`, even though used heavily during that period, is also a popular phrase and song lyric, while `life liberty and the pursuit of happiness` is a classic quote. E2R explains the popularity of these quotes through large exogeneity. The top-5 most frequent quotes of the (H,H) bucket for the E2R model are: `joe the plumber`, `I think when you spread the wealth around it's good for everybody`, `you can put lipstick on a pig`, `not god bless america god damn america`, and `the chant is drill baby drill`. In contrast with the (H,H) quotes of the E2 model *all* of these quotes were born in the 2008 U.S. Presidential elections. Thus they gained popularity not only through external media but also through the blogs network. As expected, the more powerful E2R model refines E2's anecdotal findings.

7. CONCLUSIONS

We claimed that information propagation in a social network depends not only on the dyadic relationships between

the nodes but also on the nature of the information that propagates. We characterized the nature of information items using two parameters: endogeneity and exogeneity. Given a social network and a sequence of adoptions of information items by nodes, we defined the problem of estimating the exogeneity and endogeneity parameters of the items, that best explain the observed activations. For this, we assumed that endogeneity and exogeneity are characteristics of the items that determine their propagation. We called this basic model E2. We incorporated the users' receptivity to new information in the extended model E2R. We presented efficient methods for estimating the parameters of E2 and E2R. We also showed that the estimation can be achieved with very high accuracy in synthetic scale-free graphs. Our extensive experiments with real data validated our model and yielded insightful anecdotes demonstrating the power of our model as a data-analytic tool. In the future, we plan to extend our methods for online early-prediction of exogeneity and endogeneity of items and receptivities of users. This will enhance the utility of our framework in designing the right intervention strategies for information communication.

8. REFERENCES

[1] A. Anagnostopoulos, R. Kumar, and M. Mahdian, *Influence and correlation in social networks*, KDD, 2008.

[2] S. Aral, L. Muchnik, and A. Sundararajan, *Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks*, Proceedings of the National Academy of Sciences, PNAS **106** (2009), no. 51.

[3] A.-L. Barabasi and R. Albert, *Emergence of scaling in random networks*, Science **286** (1999).

[4] N. Christakis and J. Fowler, *Connected: The surprising power of our social networks and how they shape our lives*, Back Bay Books, 2010.

[5] J. Goldenberg, B. Libai, and E. Muller, *Talk of the network: A complex systems look at the underlying process of word-of-mouth*, Marketing Letters **12** (2001), no. 3.

[6] M. Gomez-Rodriguez, J. Leskovec, and A. Krause, *Inferring networks of diffusion and influence*, 2010.

[7] A. Goyal, F. Bonchi, and L.V.S. Lakshmanan, *Learning influence probabilities in social networks*, WSDM, 2010.

[8] D. Gruhl, R. Guha, R. Liben-Nowell, and A. Tomkins, *Information diffusion through blogspace*, WWW, 2004.

[9] D. Kempe, J. Kleinberg, and É.Tardos, *Influential nodes in a diffusion model for social networks*, ICALP, 2005.

[10] D. Kempe, J. Kleinberg, and E. Tardos, *Maximizing the spread of influence through a social network*, KDD, 2003.

[11] T. Lappas, E. Terzi, D. Gunopoulos, and H. Mannila, *Finding effectors in social networks*, KDD, 2010.

[12] J. Leskovec, L. Backstrom, and J.M. Kleinberg, *Meme-tracking and the dynamics of the news cycle*, KDD, 2009.

[13] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. M. VanBriesen, and N.S. Glance, *Cost-effective outbreak detection in networks*, KDD, 2007.

[14] H. Mannila and E. Terzi, *Finding links and initiators: A graph-reconstruction problem*, SDM, 2009.

[15] M. Mathioudakis, N. Koudas, and P. Marback, *Early online identification of attention gathering items in social media*, WSDM, 2010.

[16] J.-P. Onnela and F. Reed-Tsochas, *Spontaneous emergence of social influence in online systems*, Proceedings of the National Academy of Sciences, PNAS (2010).

[17] T.J. Ypma, *Historical development of the Newton-Raphson method*, SIAM Rev. **37** (1995), no. 4.