*The Annals of Applied Statistics*
2017, Vol. 11, No. 4, 2303–2336
https://doi.org/10.1214/17-AOAS1079

# MODELING NODE INCENTIVES IN DIRECTED NETWORKS

BY DEEPAYAN CHAKRABARTI

*University of Texas, Austin*

Twitter is a popular medium for individuals to gather information and express opinions on topics of interest to them. By understanding who is interested in what topics, we can gauge the public mood, especially during periods of polarization such as elections. However, while the total volume of tweets may be huge, many people tweet rarely, and tweets are short and often noisy. Hence, directly inferring topics from tweets is both complicated and difficult to scale. Instead, the network structure of Twitter (who tweets at whom, who follows whom) can telegraph the interests of Twitter users. We propose the Producer-Consumer Model (PCM) to link latent topical interests of individuals to the directed structure of the network. A key component of PCM is the modeling of incentives of Twitter users. In particular, for a user to attract more followers and become popular, she must strive to be perceived as an expert on some topic. We use this to reduce the parameter space of PCM, greatly increasing its scalability. We apply PCM to track the evolution of Twitter topics during the Italian Elections of 2013, and also to interpret those topics using hashtags. A secondary application of PCM to a citation network of machine learning papers is also shown. Extensive simulations and experiments with large real-world datasets demonstrate the accuracy and scalability of PCM.

## 1. Introduction.

On directed social networks such as Twitter, there are two common modes of communication between individuals. A person can choose to "follow" another, so that any tweets posted by the latter are immediately routed to the former. A person can also choose to manually route their tweets by "tweeting at" another person. Both modes of directed communication express one person's belief about the preferences or interests of another. When $A$ follows $B$, $A$ clearly believes that the tweets of $B$ will be of interest to $A$. When $A$ tweets at $B$, it often implies that $A$ follows $B$ and is responding to previous tweets by $B$ on the same topic. Thus, the directed structure of the social network reflects broad topics of interest to individuals in the network.

We are interested in inferring these latent topics, and also the degree to which each individual on Twitter is interested in them. This is important for several reasons. Personalization of recommendations depends heavily on modeling user attributes such as demographics and interests [Koren (2008)]. The results of user-initiated searches can also be ranked so that results matching the user's topics of interest are on top. Such topics are also useful from a sociological perspective,

Received May 2016; revised May 2017.

*Key words and phrases.* Citation networks, social networks, stochastic blockmodel.

since they succinctly summarize the vast number of conversations between individuals into a few broad themes. Indeed, we focus on the problem of understanding Twitter topics in the polarized setting of an election campaign.

*Twitter topics during the Italian elections* (2013). We look at tweets over several months around the Italian elections of 2013 [Caldarelli et al. (2014)]. Each tweet in the dataset represents a user's comments directed at one or more users. While the text of the tweets is hidden for privacy reasons, the hashtags in the tweets are available. We seek to answer three questions. *What were the topics of interest to Italian Twitter users over this period? How did they change over the election period, if at all? Can the topics be automatically tagged with the most relevant hashtags?*

One approach would be to infer topics from the text of the tweets. However, such text can be hidden for reasons of confidentiality, as in our case. In addition, tweets are short and noisy, making text analysis difficult. The text analysis method needs to handle a large volume of tweets, and may require significant computational resources. For these reasons, topic extraction directly from the text of tweets is difficult.

We avoid these issues by inferring topics from the structure of the directed network between Twitter users. For instance, consider a person who tweets rarely. If she follows others, we can infer her interests from the topical interests of the people she follows. Computational requirements are also lower, since network structure data can be represented more concisely than tweet data. Hence, topic inference from network structure can be more scalable. Finally, the text and hashtags in tweets can be used in a post-processing step to label the inferred topics. In fact, we shall show that the same model that extracts topics from the directed network can also be used to interpret these topics using hashtags. Hence, there are strong advantages in using the network to infer topical interests of Twitter users.

The main difficulty is in relating the hidden interests of individual nodes to the observed structure of a directed network. We identify two important characteristics of directed social networks that guide the design of our model. First, the content that a person wishes to consume may differ from the content he or she produces. For instance, a wedding photographer may tweet mostly about photography, but may wish to follow sports-related personalities and journalists. Thus, the model must differentiate between the topics a person likes to tweet about (her "production" interests) and the topics she wishes to read about (her "consumption" interests). Second, Twitter makes it easy to find and follow experts on any topic. Hence, a Twitter user interested in, say, sports and music, will follow experts on each topic, rather than someone who tweets on both topics but is not an expert on either. This suggests that those who are perceived to be experts in their fields are likely to gain the most followers. Thus, for a Twitter user who wants to become popular and attract more followers, there is a strong incentive to focus on a single topic in her tweets and become known as an expert on that topic. We shall use this observation to reduce the number of parameters needed to model production interests. This model parsimony, in turn, allows our inference algorithm to scale to large graphs, where inference for more general-purpose models often fails.

It is clear that the above discussion generalizes beyond Twitter to other directed social networks such as Instagram. More interesting is its applicability to networks that are by-products of social activities but are not themselves social networks. Consider the following example.

*Topics of papers in machine learning. What are the main areas of research in machine learning? How can these areas be tagged using words used in paper titles?* We seek to answer these questions using the paper citation network, where each node is a paper, and a directed edge exists from paper $A$ to $B$ iff $A$ cites $B$. While this is not a social network, paper authors have similar incentives as Twitter users. In particular, researchers want to write papers that will be highly cited, and such papers are typically the "top" papers in one particular area of research. Thus, while an author might write papers on different topics, the content of each paper is focused on a single narrow topic (i.e., a single "production" interest). However, a paper could cite many other papers, possibly from other relevant research areas (i.e., multiple "consumption" interests). The parallels to the Twitter example are clear.

We shall show that the same model can be used for inferring topics from Twitter as well as the citation network. In addition, we shall reuse this model to label the topics with keywords as well.

1.1. *Our contributions*. We propose the Producer-Consumer Model (PCM) for directed social networks. Contributions to modeling, algorithms, and applications are discussed below.

*Model*: PCM has several important characteristics. First, it models the idea that everyone wants to be popular, and popularity on social media comes from perceived topical expertise. This incentivizes people to tweet primarily on a single topic where their expertise is greatest. To the best of our knowledge, this aspect of network structure has not been studied previously.

Second, PCM allows individuals to have multiple (but not too many) consumption interests. Consumption interests may be unrelated to their production interest. This allows for production and consumption to be handled differently in PCM.

Finally, we show that networks generated by PCM exhibit commonly observed network properties. PCM networks can match any desired in-degree distribution, including commonly observed power laws [Chakrabarti and Faloutsos (2006)]. We prove that they also exhibit reciprocity, whereby an edge from $A$ to $B$ makes the opposite edge more likely.

*Algorithm*: We present a fast and scalable algorithm to infer user interests. The algorithm alternates between finding the maximum-likelihood production interests given current estimates of consumption interests, and vice versa. We prove that each step of this procedure achieves its optimum; however, as is usual for alternating optimization methods, the final result could be only a local optimum of the overall objective. The computational complexity scales linearly with the number of topics and the maximum degree of the network, and we show empirically that it scales to large networks and is faster than competing methods.

*Accuracy and scalability*: We show the effectiveness of PCM both via simulations as well as experiments on several large citation networks from computer science and physics, and a who-trusts-whom social network. Using link prediction as a measure of model accuracy, we show that PCM is significantly more accurate than competing methods, while also scaling to large networks.

*Applications*: We show how PCM can be used not only to find topics, but also to find the most descriptive keywords (e.g., hashtags) for each topic. Thus, the topics found by PCM can be interpreted easily as well.

We use PCM to analyze the evolution of Twitter topics over the course of the 2013 Italian elections. While we expected topics to be arranged along party lines, we found this to be only partially true. Instead, recent political events, such as rallies, book releases, and current slogans, often dominate the conversation. Even seemingly party-line topics can focus on leadership struggles within parties. Our results also demonstrate the importance of TV talk shows in the Internet age. Several topics reference such talk shows, which appear to drive the conversation about politicians and parties on Twitter.

We also apply PCM to find the main research topics in machine learning from the citation network of papers. Each topic is tagged with the most relevant words from paper titles, again using PCM. We find the topics to be intuitive and reasonable.

The rest of this paper is organized as follows. We review prior work in Section 2. We present our model in Section 3. Inference of model parameters and properties of the generated networks are explored in Section 4. We present results on simulated datasets in Section 5 and on real-world datasets in Section 6. We discuss the results of PCM on the Twitter and citation networks in Section 7, followed by conclusions in Section 8. All proofs are deferred to the Supplementary Material [Chakrabarti (2017)].

## 2. Related work.

The simplest and perhaps most well-explored network model is the random graph model [Erdős and Rényi (1959), Gilbert (1959)]. Under this model, a pair of nodes (a "dyad") can be connected by an undirected edge with probability $p$, and all dyads are considered independently. The properties of this model has been widely analyzed, and extensions have been proposed to overcome deficiencies such as its Poisson degree distribution [Aiello, Chung and Lu (2000)]. The $p1$ model of Holland and Leinhardt (1981) merges the idea of dyadic independence with node-specific parameters to generate directed graphs. In particular, connections between node pairs depend on parameters reflecting expansiveness and attractiveness, which affect their out-degree and in-degree, respectively, as well as the level of reciprocity among nodes. The $p2$ model of Duijn, Snijders and Zijlstra (2004) extends this to allow for edge directionality. The $p^{\star}$ model [Frank and Strauss (1986), Wasserman and Pattison (1996)] dispenses with dyadic independence altogether; the probability of observing a particular network is modeled via an exponential family whose sufficient statistics are any set of network statistics chosen by the user. However, parameter estimation can be difficult, and is an active area of research [Caimo and Friel (2011), Hunter and Handcock (2006)].

The above models do not account for topical interests of nodes, and how they can affect the network structure. One simple approach is to extend the $p1$ model with extra latent attributes for each node, representing the topical interests of the nodes. Then linkages between nodes can be driven by these attributes. There are two popular instantiations of this idea. Hoff, Raftery and Handcock (2002) propose assigning to each node a "position" in some latent space, such that the probability of an edge between two nodes increases as the pairwise distance between them decreases. This model can capture higher-order effects such as transitivity. However, scaling such models can be difficult. Raftery et al. (2012) develop a fast inference algorithm, but the largest network they consider has only 2716 nodes. Salter-Townshend and Murphy (2013) present a variational inference method, but they only demonstrate results on a network with 604 nodes.

A second approach is the "Stochastic blockmodel," under which the probability of linkage depends solely on a node's latent cluster [Holland et al. (1983), Snijders and Nowicki (1997), Wang and Wong (1987)]: $P(i \sim j \mid \{z_u \mid u \in U\}, B) = B_{z_i,z_j}$, where $z_i$ and $z_j$ represent the latent "clusters" of node $i$ and $j$, and $U$ is the set of all nodes. The matrix $B$ represents the connection strength between clusters, with within-cluster connections (the diagonal entries of $B$) being typically higher than across-cluster connections. The number of clusters is typically picked manually.

The stochastic blockmodel has proven to be particularly fruitful, and has led to much follow-up work. The extensions mainly target two aspects of the stochastic blockmodel: edge directionality and cluster structure. Vu, Hunter and Schweinberger (2013) allow the conditional dyadic probabilities to depend on the type of dyad (reciprocal or not), its directionality, and even its sign (e.g., when edges represent positive or negative sentiments). However, it is still limited by the fact that each node must belong to a single cluster. Greater flexibility can be achieved by hierarchical clusters, or via distributions over clusters. For instance, the stochastic blockmodel leads to a block-structured matrix of linkage probabilities between nodes; this has been extended to recursive block structures such as R-MAT [Chakrabarti, Zhan and Faloutsos (2004)] and Kronecker product graphs [Leskovec et al. (2010)]. These model the idea of communities within communities, instead of a flat clustering of nodes. On the other hand, the mixed-membership stochastic blockmodels (MMSB) allows each node to have a distribution over clusters instead of belonging to a single cluster [Airoldi et al. (2008)]. The infinite relational model (IRM) extends the stochastic blockmodel by inferring the number of clusters automatically [Kemp et al. (2006)]. The idea of a latent cluster has also been extended to a larger set of latent features and attributes [Miller, Griffiths and Jordan (2009), Palla, Knowles and Ghahramani (2012), Xu et al. (2006)], to evolving graphs [Fu, Song and Xing (2009)], and to bipartite graphs [Blei, Ng and Jordan (2003) and Hofmann (1999, 2004)].

While these are related to our work, we identify some characteristics of social networks that are not easily modeled by the above approaches. First, individuals in a social network like Twitter may express their interests by following others, and by tweeting (for which they are in turn followed by others). However, one may have wide-ranging reading interests but may not necessarily tweet about the same topics. Second, none of the above models consider the incentives of the actors in the network. Scaling to large graphs is also an issue: in our experiments, none of the MCMC-based inference methods finished for the larger real-world datasets we considered. While faster approximate inference methods may scale better, they are not always available for all models. Finally, as shown later in Section 6, PCM is often more accurate than most competing models, when they did complete.

## 3. Modeling directed social networks.

Directed social networks have special characteristics that differentiate them from undirected networks. These form the basis of our proposed model.

*Separation of production and consumption.* An individual on Twitter may tweet primarily about politics, but could follow sports and music celebrities. Thus, our model must distinguish between the topic(s) of tweets written by a person from the topic(s) she is interested in reading about. We call these an individual's *production* and *consumption* interests, respectively.

*Incentives of producers.* Twitter users can easily follow experts on each topic that they are interested in consuming. Thus, if a Twitter user wants to gain many followers, the best way to do so is to be recognized as an expert on some topic. Assuming that most (if not all) individuals wish to be popular, we expect Twitter users to focus their tweets on primarily a single topic, in the hopes of being perceived as an expert on it. We model this by assuming that each user has a single production interest. This greatly reduces the number of parameters that need to be estimated, and makes the model highly scalable. Even if the assumption of a single production topic does not hold for all users, it is still a reasonable approximation that yields significant benefits in scalability.

*Multiple consumption topics per person.* We allow each individual to have different degrees of interest in the topics they wish to consume. We represent this as an unnormalized vector of topical interests for each individual. This is different from modeling consumption interests as a (normalized) distribution over topics, where information about the total degree of interest is lost. Also, each individual has limited attention and cannot be interested in too many topics. We model this by a prior over the vectors mentioned above.

*Unlinking topic associations and node popularity.* A generative network model must account for two very different aspects of the nodes: their *popularity* (or, degree), and their *topic associations*. Some models, such as the Stochastic Blockmodel, try to capture both aspects using the set of parameters. In such cases, matching node degrees might acquire outsized importance in model-fitting computations, hurting the inference of the latent topics. We focus solely on topic inference, which we believe is more important in real-world applications such as recommendation systems, where matching user and product topics is of primary importance. Hence, our generative model uses separate parameters for node popularity and topics. This is similar in spirit to degree-corrected stochastic blockmodels, which separate node degrees and latent topics [Karrer and Newman (2011)].

3.1. *Model specification.* Let $A$ be the adjacency matrix representing a directed graph, with $A_{uv} = 1$ if a directed edge exists from $u$ to $v$, and $A_{uv} = 0$ otherwise. Rows and columns of $A$ represent consumers and producers, respectively. Note that every Twitter user is both a producer and a consumer, and is represented as both a row and a column. However, the model allows arbitrary sets of producers and consumers. We denote the set of consumers and producers by $U$ and $V$, respectively, with $N = |U|$ being the number of consumers, and $I_v = \{u \in U \mid A_{uv} = 1\}$ being the set of followers of $v \in V$ (the *in-links* of $v$). Let $\mathcal{T}$ represent the set of topics, with $K = |\mathcal{T}|$. The model has four sets of parameters that capture the characteristics of directed networks mentioned earlier.

*Consumption interests*: The probability that a person $u \in U$ is interested in consuming tweets on topic $t \in \mathcal{T}$ is denoted by $\theta_{ut}$. Thus, the set $\{\theta_{ut} \mid t \in \mathcal{T}\}$ represent all the consumption interests of $u$. Note that we do not require that $\sum_t \theta_{ut} = 1$.

*Production interests*: A content producer on Twitter is incentivized to tweet on a single topic, to enhance the perception that she is an expert on that topic, and hence gain followers interested in consuming that topic. For each producer $v \in V$, we use $t_v \in \mathcal{T}$ to denote this single topic on which $v$ tweets.

*Popularity*: Two producers who tweet on the same topic may still have differences in the quality of their tweets, and hence different popularities. The popularity of a producer $v$ is represented by a parameter $n_v$. We shall present its precise definition shortly, but intuitively, for large $N$, this is the number of followers of $v$.

*Purity*: Finally, for each $v \in V$, we shall need a parameter $\alpha_v \in \mathbb{R}_{>0}$ to represent the degree to which the followers of $v$ are actually interested in the topic $t_v$ of her tweets, as against their following $v$ due to other extraneous factors. For instance, a celebrity who tweets about music may have some followers who are interested in music, and also other followers who simply wish to follow celebrities. Another example is that of people following others simply because they are friends in real life, and not because of any particular shared interest. High values of $\alpha_v$ will mean that topical match is the driving factor behind whether one follows $v$ or not. Hence, we will refer to $\alpha_v$ as the "purity" of $v$. We shall require $\alpha_v < (N - N_{t_v})/N_{t_v}$, where $N_t = \sum_u \theta_{ut}$ denotes the total "weight" of topic $t$.

Given the model parameters $(\{\theta_{ut} \mid u \in U, t \in \mathcal{T}\}, \{t_v, n_v, \alpha_v \mid v \in V\})$, the network is created by independently generating the $n_v$ followers for each producer $v$. The followers are drawn from a multinomial distribution over the set of consumers $U$, where the multinomial probability $\eta_v(u)$ that $u$ follows $v$ depends on the consumption interest $\theta_{u,t_v}$ of $u$ in the production interest $t_v$ of $v$. Specifically, let

$$\eta_v(u) = \theta_{u,t_v} \cdot p_v + (1 - \theta_{u,t_v}) \cdot q_v, \tag{3.1}$$

where

$$p_v = (1 + \alpha_v)/N, \tag{3.2}$$

$$q_v = \bigl(1 - \alpha_v \cdot N_t/(N - N_{t_v})\bigr)/N. \tag{3.3}$$

Then draw $n_v$ samples with replacement from the multinomial distribution given by $\{\eta_v(u) \mid u \in U\}$ [note that $\sum_u \eta_v(u) = 1$ by construction]. Let $s_v(u)$ denote the number of times $u$ is drawn among these $n_v$ samples. These samples, after duplicate removal, become the followers of $v$. In other words, $A_{uv} = I\{s_v(u) \geq 1\}$. This process is repeated independently for each producer to generate the entire network. We call this model the Producer-Consumer Model (PCM).

The intuition is as follows. Consider a consumer $u$ who is extremely interested in a topic $t$, and producer $v$ tweets on that very topic, that is, $t_v = t$ and $\theta_{u,t_v} = 1$. Then the multinomial probability of selecting $u$ in one draw is $\eta_v(u) = p_v = (1 + \alpha_v)/N$. In other words, the chances of selecting $u$ in one draw are elevated by $\alpha_v/N$ over the uniform distribution. On the other hand, if $u$ had no interest in $t_v$ ($\theta_{u,t_v} = 0$), then the chances of selecting $u$ in one draw would be $\eta_v(u) = q_v < 1/N$. Note that $\theta_{ut} = 1$ does not imply that $u$ follows every producer of topic $t$; even among the latter, there will be differences in the quality of tweets, which will be reflected in the number of followers for these producers. The node purity parameter $\alpha_v$ controls the degree to which followers of $v$ are actually interested in $t_v$. At its maximum value of $\alpha_v = (N - N_{t_v})/N_{t_v}$, we have $q_v = 0$, and interest match is the sole reason for following $v$.

The use of latent interests combined with directionality has been explored in the literature [Duijn, Snijders and Zijlstra (2004), Fosdick and Hoff (2015), Hoff (2005), Krivitsky et al. (2009), Vu, Hunter and Schweinberger (2013)]. However, to the best of our knowledge, the ideas of people following experts, and of producers having different "purities," have not been explored in prior work. For instance, consider the popular Mixed Membership Stochastic Blockmodel (MMSB), which sets $P(A_{uv} = 1 \mid \{z_i\}, B) = z_u^t B z_v$ for some multinomial distribution $z_i$ of interests for each node $i$, and a matrix $B \in [0, 1]^{K \times K}$. Both MMSB and PCM use latent topic vectors for each node. However, MMSB does not differentiate between the production and consumption interests of a node (though other models do so, as mentioned above). Also, the probability of $u$ following $v$ is based on measuring similarity over all topics, weighted by the matrix $B$. In contrast, PCM models the idea that consumers follow experts on each topic. In the MMSB notation, this would roughly correspond to the requirement that $P(A_{uv} = 1 \mid \{z_i\})$ be an increasing function of, say, $\sum_t z_u(t) \cdot I\{z_v(t) > \tau\}$, for some threshold $\tau$ that is close to 1. We believe that PCM achieves this effect more naturally. Finally, the node purity parameter of PCM has no direct counterpart in MMSB.

A related model [Hoff (2009)] sets log odds $(P(A_{uv} = 1 \mid \boldsymbol{\beta}, \mathbf{Z}, \mathbf{B}, \mathbf{W}, \mathbf{E})) = \mathbf{x}_{uv}^t \boldsymbol{\beta} + \mathbf{z}_u^t B \mathbf{w}_v + \epsilon_{uv}$, where $\mathbf{x}_{uv}$ represents features of the ordered dyad $(u, v)$, $\mathbf{Z}$ and $\mathbf{W}$ represent latent features of nodes, and $\mathbf{E}$ is a matrix of standard normal noise. This model does differentiate between producers and consumers. However, like the Mixed Membership Stochastic Blockmodel, the similarity of topical interests is computed via a weighted dot-product $(\mathbf{z}_u^t B \mathbf{W})$, which does not model the idea of following experts. There is also no notion of node purity.

These differences also hold between PCM and latent distance models. For instance, ignoring covariates, the model of Hoff, Raftery and Handcock (2002) sets $P(A_{uv} = 1 \mid \{z_i\}) = f(|z_u - z_v|)$, where $z_i$ is the latent node positions of node $i$, and $f(\cdot)$ is a monotonically decreasing function. In such a model, the notion of "expertise" on a given topic is unclear. Also, there is no concept of node purity. Indeed, the probability of a link from $u$ to $v$ in PCM depends not only on the topical match (like the $|z_u - z_v|$ term) but also on node $v$ itself (via its purity $\alpha_v$); a single $f(\cdot)$ for all nodes cannot model this.

*Likelihood*: To write down the likelihood of PCM, we first define the set $Q_v(n_v)$ of all possible multinomial draws that, after duplicate removal, yield the observed set of followers $I_v$ of $v$. Letting $\mathbb{W}$ represent the set of whole numbers, we find

$$Q_v(n_v) = \{w \in \mathbb{W}^{|I_v|} \mid w(u) = 0\ \forall u \notin I_v, w(u) \geq 1\ \forall u \in I_v, |w|_1 = n_v\}. \tag{3.4}$$

$Q_v(n_v)$ is defined to be the empty set if $n_v < |I_v|$. Then the likelihood of the parameters $(\{\theta_{ut}\}, \{t_v, n_v, \alpha_v\})$ is given by

$$\begin{aligned} &L(\{\theta_{ut}\}, \{t_v, n_v, \alpha_v\} \mid A) \\ &\quad = \prod_{v \in V} \left[ I\{n_v \geq |I_v|\} \cdot n_v! \cdot \sum_{w \in Q_v(n_v)} \prod_{u \in I_v} \frac{\eta_v(u)^{w(u)}}{w(u)!} \right]. \end{aligned} \tag{3.5}$$

*Prior*: As it stands, the model allows a user to be extremely interested in all topics ($\theta_{ut} = 1$ for all $t$). Clearly, this is unlikely. Hence, we add a prior to express our belief that consumers tend to be interested in only a few topics. We will model this as two constraints on the total consumption interest $\sum_t \theta_{ut}$ for any consumer $u$. First, we place a hard constraint: $\sum_t \theta_{ut} \leq \tau$ for some $\tau > 1$. Second, we place a penalty that increases as $\sum_t \theta_{ut}$ grows greater than 1. More precisely, we set

$$\begin{aligned} &P(\{\theta_{ut} \mid u \in U, t \in \mathcal{T}\}) \\ &\quad = \prod_{u \in U} \left[ \frac{1}{Z} \cdot I\left\{ 0 \leq \theta_{ut} \leq 1\ \forall t, \sum_t \theta_{ut} \leq \tau \right\} \cdot e^{-\lambda \cdot \max\{0, \sum_t \theta_{ut} - 1\}} \right]. \end{aligned} \tag{3.6}$$

The constant of proportionality $Z$ will not be needed henceforth, but the Supplementary Material [Chakrabarti (2017)] (Proposition A.5) provides a formula for when $\tau$ is integral.

Now, combining equations (3.5) and (3.6) and ignoring constants, we can write the posterior as

$$\begin{aligned} &P(\{\theta_{ut}\}, \{t_v, n_v, \alpha_v\} \mid A) \\ &\quad \propto \prod_{v \in V} \left[ I\{n_v \geq |I_v|\} \cdot n_v! \cdot \sum_{w \in Q_v(n_v)} \prod_{u \in I_v} \frac{\eta_v(u)^{w(u)}}{w(u)!} \right] \\ &\qquad \times \prod_{u \in U} \left[ I\left\{0 \leq \theta_{ut} \leq 1 \ \forall t, \sum_t \theta_{ut} \leq \tau \right\} \cdot e^{-\lambda \cdot \max\{0, \sum_t \theta_{ut} - 1\}} \right]. \end{aligned} \tag{3.7}$$

*Simplifying the posterior.* The above formula can be simplified via the next proposition.

PROPOSITION 3.1. *If $n_v < \sqrt{\frac{N}{1+\alpha_v}}$, the MAP estimate of $\eta_v$ is $|I_v|$.*

We shall henceforth assume that the condition of Proposition 3.1 is true (see remarks below). Thus, the MAP estimate of $\eta_v$ is simply $|I_v|$, for any producer $v$. After plugging this into the log-posterior, parameter inference reduces to solving the following optimization:

$$\begin{aligned} &\text{Maximize} \sum_{v \in V} \sum_{u \in I_v} \log(\theta_{u,t_v} \cdot p_v + (1 - \theta_{u,t_v}) \cdot q_v) \\ &\qquad - \lambda \sum_{u \in U} \max\left\{0, \sum_t \theta_{ut} - 1\right\} \\ &\text{subject to} \sum_t \theta_{ut} \leq \tau \qquad \forall u \in U \\ &\qquad \text{and} \quad 0 \leq \theta_{ut} \leq 1 \qquad \forall u \in U \text{ and } t \in \mathcal{T}, \end{aligned} \tag{3.8}$$

where $p_v$ and $q_v$ are defined in equations (3.2) and (3.3).

REMARK 1. There are two reasons for assuming the condition of Proposition 3.1. First, the in-degree distribution of many networks follow heavy-tailed distributions [Chakrabarti and Faloutsos (2006), Handcock and Jones (2004)], so the bulk of producers will satisfy this condition. Second, we expect individuals with extremely high in-degree to have many followers who are merely interested in following celebrities (we will show this later in Figure 4). Thus, their purity will be low ($\alpha_v \approx 0$, so $p_v \approx q_v$), which implies that their contribution to the likelihood will be nearly constant. Hence, we can safely remove any nodes with extremely high in-degrees in a preprocessing step. We note that such removal has been shown to be useful in graph search as well [Sarkar and Moore (2010)].

**4. Analysis and inference.** In this section, we will find the MAP estimates $\hat{\theta}_{ut}$, $\hat{t}_v$, and $\hat{\alpha}_v$. Then we will discuss related issues such as identifiability and the properties of networks generated by PCM.

*Node purity* ($\alpha_v$). The MAP estimate of $\alpha_v$ is given by the root of a function, but has no closed form.

THEOREM 4.1. *The MAP estimate of $\alpha_v$ is given by $\hat{\alpha}_v = \max(0, y)$, where $y$ satisfies*

$$\sum_{u \in I_v} \frac{1}{1/r_{u,\hat{t}_v} + y} = 0 \qquad \textit{with } r_{ut} = \frac{\hat{\theta}_{ut} - N_t/N}{1 - N_t/N}, \tag{4.1}$$

*where $\hat{\theta}_{ut}$ and $\hat{t}_v$ are the MAP estimates of $\theta_{ut}$ and $t_v$, respectively, and $N_t = \sum_u \hat{\theta}_{ut}$.*

The left-hand side of equation (4.1) is a monotonic function of $y$, so its solution can be found quickly via binary search.

To gain intuition, consider the term $f(\hat{\theta}_{ut}) \triangleq (1/r_{ut} + y)^{-1}$ as a function of $\hat{\theta}_{ut}$ for a fixed $y$. This can be shown to be a concave function of $\hat{\theta}_{ut}$, and which is nearly linear for small $y$. This suggests approximating the function by a straight line between its end-points ($\hat{\theta}_{ut} = 0$ and 1), yielding

$$f(x) \approx f(0) + x\big(f(1) - f(0)\big) = \frac{1}{N/N_t - 1 - y}\left(-1 + x \cdot \frac{N/N_t}{(1+y)}\right).$$

Using this approximation in equation (4.1) yields

$$\hat{\alpha}_v = \max\left\{0, \frac{\sum_{u \in I_v} \theta_{u,\hat{t}_v}/n_v}{\sum_{u \in U} \theta_{u,\hat{t}_v}/N} - 1\right\}.$$

Thus, $\hat{\alpha}_v$ can be interpreted as the average interest of the followers of $v$ in the topic on which $v$ tweets, normalized against a baseline popularity of that topic among all consumers. This matches our intuition that, for a "pure" producer $v$, a large fraction of her followers are actually interested in consuming topic $t_v$.

*Producer topic* ($t_v$). The production interest $t_v$ of a producer $v$ can be computed from the consumption interests of her followers $I_v$.

PROPOSITION 4.2. *The MAP estimate of $t_v$ is given by*

$$\hat{t}_v = \arg\max_t \prod_{vt} \qquad \textit{with } \prod_{vt} = \prod_{u \in I_v} \big(\hat{\theta}_{ut} \cdot \hat{p}_v + (1 - \hat{\theta}_{ut}) \cdot \hat{q}_v\big),$$

*where $\hat{\theta}_{ut}$ is the MAP estimate of $\theta_{ut}$, and $\hat{p}_v$ and $\hat{q}_v$ are obtained from equations* (3.2) *and* (3.3) *with the MAP estimate $\hat{\alpha}_v$ plugged in for $\alpha_v$.*

Computationally, both $\hat{\alpha}_v$ and $\hat{t}_v$ can be computed from the parameters of the followers of $v$, and this scales linearly with the number of followers and the number of topics. The number of followers (i.e., the in-degrees) of the producer nodes are often heavy-tailed, with most nodes having low degrees. In addition, the maximum in-degree is limited (see Remark 1). Hence, fitting these parameters is fast and can scale to large real-world networks.

*Consumer interests* ($\theta_{ut}$). Given $\hat{\alpha}_v$ and $\hat{t}_v$, the problem of finding the MAP estimate of $\boldsymbol{\theta_u} = \{\theta_{ut} \mid t \in \mathcal{T}\}$ corresponds to

$$\begin{aligned} &\text{maximizing } \sum_{v \in O_u} \log(\theta_{u,\hat{t}_v} \cdot \hat{p}_v + (1 - \theta_{u,\hat{t}_v}) \cdot \hat{q}_v) - \lambda \max\left\{0, \sum_t \theta_{ut} - 1\right\} \\ &\text{subject to } \sum_t \theta_{ut} \le \tau \quad \text{and} \quad 0 \le \theta_{ut} \le 1, \end{aligned} \tag{4.2}$$

where $O_u = \{v \in V \mid A_{uv} = 1\}$ is the set of producers who are followed by $u$ (i.e., the *out-links* of $u$).

Now, we state our algorithm to infer $\boldsymbol{\theta_u}$. Define $\kappa_v = \hat{q}_v/(\hat{p}_v - \hat{q}_v)$. Note that $\kappa_v > 0$ since $\hat{p}_v > \hat{q}_v$. Define $\Gamma_u = \bigcup_{v \in O_u}\{\hat{t}_v\}$ to be the set of topics of the producers $O_u$ followed by $u$. For each pair $(u, t)$ such that $t \in \Gamma_u$, define $y_{ut} = \min\{\kappa_v \mid v \in O_u, \hat{t}_v = t\}$ and the function $h_{ut} : (-y_{ut}, \infty) \to \mathbb{R}^+$ as $h_{ut}(x) = \sum_{v \in O_u} I\{\hat{t}_v = t\}/(x + \kappa_v)$. Note that $h_{ut}(x)$ is a monotonically decreasing bijective function, and its inverse exists. Define $w_{ut}(\ell) = \max(\min(h_{ut}^{-1}(\ell), 1), 0)$; this takes the inverse of $h_{ut}$ and clips the result to between 0 and 1 (the allowable range of $\theta_{ut}$).

THEOREM 4.3. *Given $\{\hat{\alpha}_v, \hat{t}_v\}$, Algorithm 1 finds the MAP estimates $\{\hat{\theta}_{ut}\}$ for any $u \in U$.*

*Summary of inference algorithm.* Starting from a random initialization, we iterate between producers and consumers until convergence. Given consumer interests $\hat{\theta}_{ut}$, we infer the unique production interest $\hat{t}_v$ and the node purity $\hat{\alpha}_v$ of each producer using Proposition 4.2 and Theorem 4.1, respectively. Then, armed with the producer-specific parameters, the interests $\hat{\theta}_{ut}$ of the consumers on various topics values are updated via Algorithm 1. Convergence is guaranteed since each step increases the value of the objective (Proposition 4.2 and Theorem 4.3); however, it might be a local optimum.

*Combining production and consumption interests.* We have so far kept production and consumption interests separate, even when they are for the same Twitter user. This is appropriate when a Twitter user has many in-links as well as out-links, since there is enough information to infer her production and consumption interests separately. However, inference can be more difficult in sparser settings.

**Algorithm 1** Inferring $\boldsymbol{\theta_u}$

```
Require: Set $O_u$ of producers followed by $u$; MAP estimates $\{\hat{\alpha}_v, \hat{t}_v\}$ for all producers $v \in O_u$
if $\sum_t w_{ut}(\lambda) < 1$ then
    Find $\ell \in [\min_t h_{ut}(1), \lambda)$ such that $\sum_t w_{ut}(\ell) = 1$ via binary search.
else if $1 \leq \sum_t w_{ut}(\lambda) \leq \tau$ then
    $\ell = \lambda$.
else if $\sum_t w_{ut}(\lambda) > \tau$ then
    Find $\ell \in (\lambda, \max_t h_{ut}(0)]$ such that $\sum_t w_{ut}(\ell) = \tau$ via binary search.
end if
$\hat{\theta}_{ut} = w_{ut}(\ell)$
return $\{\hat{\theta}_{ut}\}$
```

In particular, estimates of production interests can be noisy for Twitter users who have few followers.

For such cases, we shall assume that the consumption interests of an individual also reflect her production interest. In particular, the chance that $v$ has production interest $t_v$ is proportional to $\theta_{u,t_v}$. Thus, the contribution to the log-posterior [equation (3.8)] for each producer $v$ now becomes

$$\log\left(\sum_t \frac{\theta_{vt}}{\sum_{t'} \theta_{vt'}} \prod_{u \in I_v} (\theta_{ut} \cdot p_v + (1 - \theta_{ut}) \cdot q_v)\right), \tag{4.3}$$

instead of $\log(\prod_{u \in I_v} (\theta_{u,t_v} \cdot p_v + (1 - \theta_{u,t_v}) \cdot q_v))$.

Note that this posterior still differentiates between production and consumption. This is because the normalized value $\theta_{vt} / \sum_{t'} \theta_{vt'}$ is used in determining the production interest of $v$, while the unnormalized $\theta_{vt}$ represent her consumption interests. For inference, proximal gradient descent methods initialized with the results of Proposition 4.2 and Algorithm 1 can be used.

*Parameter selection.* PCM has three user-specified parameters: the number of topics $K$, the hard threshold $\tau$ on the total consumption interest for any consumer, and a corresponding soft penalty parameter $\lambda$. We can select all three automatically, via a link-prediction task. Specifically, we randomly remove some fraction of links from the given network, train PCM on the remaining network, and then use the inferred parameters to predict the edges that were most likely to have been removed. We compare the accuracy of this link-prediction task for a range of $(K, \tau, \lambda)$ tuples, replicating each experiment multiple times with different edges being removed each time. The tuple that gives the best link-prediction accuracy can now be used to infer PCM parameters over the entire network. When the desired number of topic $K$ is fixed by the user, the other two parameters can be chosen in this fashion. Empirical results shown later in Section 6 validate this approach.

*Identifiability.* Consider two parameter settings $(\{\theta_{ut}\}, \{t_v, \alpha_v\})$ and $(\{\theta'_{ut}\}, \{t'_v, \alpha'_v\})$. Both yield the same likelihood for any generated network if $\theta_{u,t_v} \cdot p_v + (1 - \theta_{u,t_v}) \cdot q_v = \theta'_{ut'_v} \cdot p'_v + (1 - \theta'_{ut'_v}) \cdot q'_v$ for all $(u, v)$. The conditions under which this happens is given by the following theorem.

THEOREM 4.4 (Identifiability). *Given two feasible parameter settings* $(\{\theta_{ut}\}, \{t_v, \alpha_v\})$ *and* $(\{\theta'_{ut}\}, \{t'_v, \alpha'_v\})$, *we have* $\theta_{u,t_v} \cdot p_v + (1 - \theta_{u,t_v}) \cdot q_v = \theta'_{ut'_v} \cdot p'_v + (1 - \theta'_{ut'_v}) \cdot q'_v$ *for all* $(u, v)$ *iff there exist* $\{a_t, b_t\}$ *such that*

$$\theta'_{ut} = a_t \cdot \theta_{ut} + b_t,$$

$$\alpha'_v = \frac{\alpha_v}{a_t} \cdot \left(\frac{1 - b_t - a_t \cdot N_t/N}{1 - N_t/N}\right).$$

Thus, the model is not identifiable if $\theta'_{ut}$ is a feasible scaled and translated version of $\theta_{ut}$, that is, $\theta'_{ut} = a_t \cdot \theta_{ut} + b_t$ for some feasible $(a_t, b_t)$. However, it is identifiable up to a permutation of topics under the following condition.

COROLLARY 4.5. *There exists at most one solution* (*up to permutation of topics*) *where, for each topic* $t$, *the sets* $\{u \in U \mid \theta_{ut} = 1\}$ *and* $\{u \in U \mid \theta_{ut} = 0\}$ *are nonempty.*

The proof follows from observing that only $a_t = 1$ and $b_t = 0$ satisfies the conditions of Theorem 4.4 while ensuring $\alpha'_v \geq 0$. A sufficient condition for Corollary 4.5 is that there must be a "pure node" for each topic, that is, a consumer who is extremely interested in that topic, and in nothing else. This mirrors conditions used to prove consistency in stochastic blockmodel variants [see, for instance, Zhang, Levina and Zhu (2014)].

Using this identifiability condition, we can prove the existence of a MAP solution for the posterior of equation (3.8).

THEOREM 4.6 (Existence of MAP). *Under the conditions of Corollary* 4.5, *the MAP solution exists.*

### 4.1. *Network properties.*

Several patterns observed in real-world networks are exhibited by PCM networks. We look at two important properties: degree distributions, and reciprocity.

*Degree distributions.* Any desired in-degree distribution can be modeled by placing the corresponding prior on the parameters $n_v$ representing the number of consumers interested in producer $v$. Note that a node's popularity $(n_v)$ is separate from her production interest $(t_v)$ or her consumption interests $(\theta_{ut})$, and this is an important aspect of PCM.

*Reciprocity.* This refers to the phenomenon of a person $v$, on being followed by another person $u$, "returning the favor" by following $u$ in turn. This can be expressed as increased chances of the "reciprocal" link, i.e., $P(A_{vu} = 1 \mid A_{uv} = 1) \geq P(A_{vu} = 1)$. Such a relationship can be shown to exist under PCM as well.

Suppose each node is both a producer and a consumer. First, let us consider the case where interests are binary: $\theta_{ut} \in \{0, 1\}$ for all $(u, t)$. Then, the next theorem shows that reciprocity is guaranteed if the production interest $t_v$ of every node $v$ is also one of her consumption interests (e.g., someone who tweets about "nature" photographs will also be interested in following other "nature" photographers).

THEOREM 4.7 [Reciprocity (binary interests)]. *Suppose all nodes are both producers and consumers. Let $C_{t_v} = \{u \mid \theta_{u,t_v} = 1\}$ be the set of consumers interested in topic $t_v$ of tweets written by $v$. If $v \in C_{t_v}$ for all nodes $v$, then $P(A_{vu} = 1 \mid A_{uv} = 1) \geq P(A_{vu} = 1)$.*

For intuition, consider an edge from $u$ to $v$. This suggests that $u$ is probably interested in consuming $t_v$ (the production interest of $v$). But $u$ is interested in consuming $t_u$ as well, by the condition of the theorem. This implies a greater than random chance that $t_u = t_v$, with the chances being higher if people have few consumption interests. Since $v$ is interested in $t_v$ (again, by the theorem's condition), we find that $v$ has an elevated chance of being interested in $t_u$, and hence being a follower of $u$. Thus, the model exhibits reciprocity.

In the general case, $\theta_{ut} \in [0, 1]$. Following the earlier intuition, we may expect reciprocity if a greater consumption interest in topic $t$ implies greater chances of $t$ being the production topic of that user as well. Formally, let $T = \{t_v\}$ be drawn from some distribution instead of being fixed parameters. Let $\Theta = \{\theta_{ut} \mid u \in U, t \in \mathcal{T}\}$.

THEOREM 4.8 [Reciprocity (general case)]. *Suppose all nodes are both producers and consumers. If $P(T \mid \Theta) = \prod_u g_u(\theta_{u,t_u})$ for some monotonically nondecreasing functions $g_u$, then $P(A_{vu} = 1 \mid A_{uv} = 1) \geq P(A_{vu} = 1)$.*

## 5. Simulations.

In this section, we will test the inference procedure for PCM. We simulate directed networks generated by PCM with varying number of topics, node degrees, node purities, and number of consumption interests. Then we recover these topics from the networks, and compare our inferences to the ground truth.

*Setup.* We generate graphs with $N = 300$ nodes, and 2 to 5 topics. The in-degrees of nodes are drawn from a power-law distribution to match the commonly observed heavy-tailed nature of degree distributions [Chakrabarti and Faloutsos (2006)]:

$$P(\text{node } i \text{ has degree } d_i) \propto d_i^{-\gamma_{\text{deg}}}.$$

We vary the parameter $\gamma_{\text{deg}}$ in our simulations, with larger values implying greater probability of low degrees. We set the minimum degree to 4, since inference of production interests with too few followers is unrealistic.

We assign binary consumption interests to each node ($\theta_{ut} \in \{0, 1\}$) as follows. First, the number of interests of a node is picked via a power-law with parameter $\gamma_{\text{int}}$. This ensures that most nodes exhibit only a few interests, but there are also some nodes have many interests. Then these many consumption interests are selected uniformly from among all possible topics. Once the consumption interests of a node are drawn, its production interest is selected uniformly at random from among its consumption interests. This follows the intuition that people tweet about topics they care about, and hence are also interested in reading others' tweets on the same topic. This also matches the reciprocity condition outlined in Theorem 4.8.

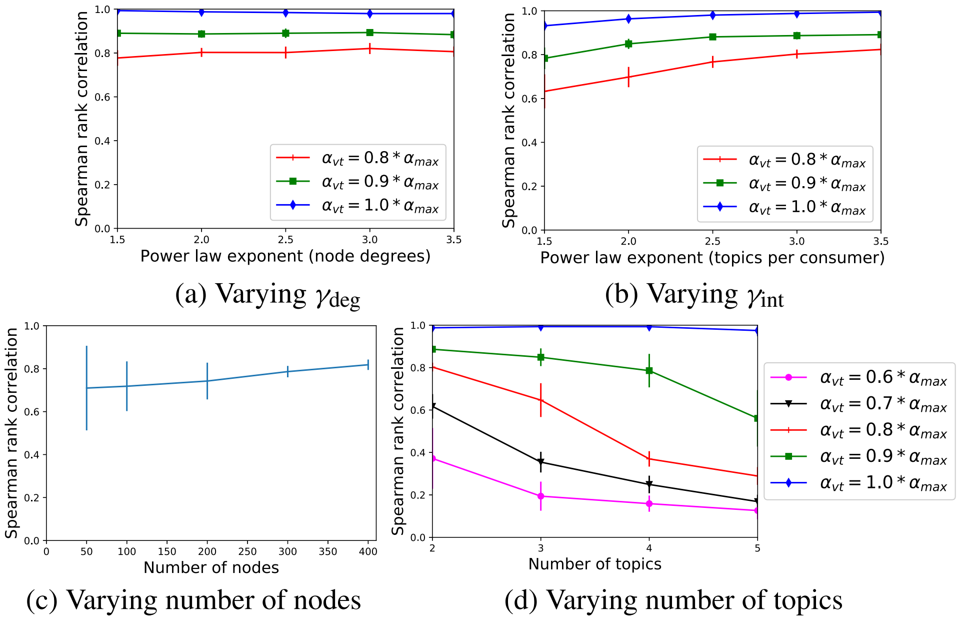
*Evaluation.* For each topic, we first rank nodes in order of their estimated consumption interest for that topic. Then we measure the Spearman rank correlation between this ranking and the ground truth ranking based on the true consumption interests. We report the average Spearman correlation over all topics. Since the estimated topics can be a permutation of the ground truth topics, we search for the best-fit permutation before computing the above measure.

We note that the Spearman rank correlation is a particularly stringent measure of accuracy, for two reasons. First, the ground truth ranking has ties (since true consumption interests are binary), but inferred interests will rarely have ties. This mismatch hurts the rank correlation. Second, for applications such as link prediction or recommendations, we only need to identify people whose interest in a topic is above a given threshold. The precise ranking of individuals is unnecessary for this purpose. We choose to report the Spearman rank coefficient only because it magnifies the differences between parameter settings, and makes trends obvious.

*Results.* Figure 1(a) shows the dependence of rank correlation on node in-degrees ($\gamma_{\text{int}}$ is set to 3.0). As $\gamma_{\text{deg}}$ increases, smaller degrees become more likely. However, the rank correlation measure is unaffected, showing that PCM is robust to the degree distribution of the network.

Figure 1(b) shows the dependence of rank correlation on the distribution of the number of consumption interests of nodes ($\gamma_{\text{deg}}$ is set to 2.0). Increasing $\gamma_{\text{int}}$ leads to fewer interests on average. We see that rank correlation improves when consumers have fewer interests on average. This is because a consumer with few interests makes it easier to disambiguate the production interests of everyone she follows. Better estimates of production interests in turn leads to improved estimates of consumption interests.

Figure 1(c) shows how the rank correlation varies with the network size. The networks were generated using $\gamma_{\text{int}} = 3$, $\gamma_{\text{deg}} = 3$, and $K = 2$ topics. Inference is more accurate for larger networks, while smaller networks show low average rank correlation and greater variations in accuracy.



(a) Varying $\gamma_{\text{deg}}$

(b) Varying $\gamma_{\text{int}}$

(c) Varying number of nodes

(d) Varying number of topics

FIG. 1. *Spearman rank correlation as a function of simulation parameters.*

Figure 1(d) shows how the rank correlation varies with number of topics. We also vary the node purity, which we report as a fraction of the maximum possible purity $\alpha_{\max} = (N - N_t)/N_t$. The average Spearman correlation increases with fewer topics and purer producers, as expected.

In summary, we find that inference under PCM is robust to changes in the distribution of node degrees, and performs the best when there are few topics, purer nodes, and fewer consumption interests. The number of topics $K$ has the greatest impact on performance. This is expected, since the number of parameters depends linearly on $K$. We note that our alternative minimization procedure for parameter inference can get stuck in local minima, and may not actually find the MAP estimate. However, the simulations show that inference is reasonable, especially for high node purities. Indeed, our inference yields better results in link-prediction tasks than competing methods (to be discussed next in Section 6). Additionally, we shall show that inferred node purities are indeed high in our applications (Section 7), suggesting that PCM inference is appropriate there.

## 6. Experiments with real datasets.

We now turn to experiments with real-world datasets. Since the ground truth is unknown, we will judge the accuracy of PCM using a link prediction task. We will show that PCM is more accurate than competing methods, suggesting that it finds better topics. These experiments will also showcase the scalability of PCM. We will also demonstrate its robustness with respect to parameter settings.

*Data.* We report results on multiple citation and social network datasets. As noted earlier, citation networks reflect the underlying incentives of the authors; a paper may cite multiple papers in its own field as well as related fields to demonstrate broad scope, yet the paper may contribute to, and be known as an exemplar of, its primary topic. Our datasets consist of citation networks in (a) Machine Learning (ML), (b) Robotics/AI (AI/ROBOTICS), (c) Computer Science Theory (THEORY), (d) High-energy physics theory (HEP-TH), and (e) High-energy physics Phenomenology (HEP-PH). The first three are derived from Citeseer [Caragea et al. (2014)], and the next two from Gehrke, Ginsparg and Kleinberg (2003). We also use (f) the Epinions who-trusts-whom social network (EPINIONS) [Richardson, Agrawal and Domingos (2003)], where one user may trust another primarily for her reviews and ratings of one product type (say, the electronics "topic"). Network statistics are presented in Table 1.

*Models.* In addition to PCM, we ran experiments with three popular methods: the Mixed Membership Stochastic Blockmodel (MMSB) [Airoldi et al. (2008)], the Infinite Relational Model (IRM) [Kemp et al. (2006)], and the SVI variant of MMSB [Gopalan and Blei (2013)]. IRM assigns each node to a latent cluster and predicts the probability of an edge from $u$ to $v$ based on the clusters of the corresponding nodes (i.e., the "stochastic blockmodel" approximation); IRM selects the number of blocks automatically. MMSB also uses a block model, but it associates a distribution over topics for each node, and predicts the link from $u$ to $v$ by drawing from the topic vectors of $u$ and $v$. We used the MCMC-based inference method of Chang (2012). SVI is a fast variational inference method for MMSB that has been used in finding overlapping communities. We could not compare against the Infinite Latent Attribute model (ILA) [Palla, Knowles and Ghahramani (2012)] since it failed with even our smallest dataset.

*Evaluation via link prediction.* Since the actual node clusters or topics are unknown in our datasets, the accuracy of the models must be measured indirectly. To gauge model quality, we compare the accuracies of the various models on a link prediction task, which is a common approach for evaluating such models. In particular, 10% of directed edges are selected at random from the full network to create a test set $E^{(\text{test})}$, while the remaining links form the training network $A^{(\text{train})}$. Each model is trained on $A^{(\text{train})}$, and then required to predict the missing followers of each producer in $E^{(\text{test})}$, that is, the missing followers of $V^{(\text{test})} = \{v \mid u \to v \in E^{(\text{test})}\}$.

More precisely, for each $v \in V^{(\text{test})}$, let $S = \{u \mid A^{(\text{train})}_{uv} = 0\}$ be the set of nodes who do not follow $v$ in $A^{(\text{train})}$. For each $u \in S$, consider the augmented network $A'(u)$ that is identical to $A^{(\text{train})}$ except it has an extra link from $u$ to $v$. We measure the probability of observing $A'(u)$ using the parameters $\{\theta_{ut}, t_v, \alpha_v\}$ inferred from $A^{(\text{train})}$. We then order all nodes $u \in S$ in decreasing order of the probability

TABLE 1

*Average AUC of link prediction*: (a) *PCM outperforms other models when $K = 10$ topics are allowed* (*except IRM*, *which picks $K$ automatically*). *The closest competitor is SVI*, *which performs better for* Hep-Ph. (b) *When the optimal parameter settings are used*, *PCM always outperforms SVI. The "×" symbol is used when the method failed for the given dataset*; *for example*, *the size of* Hep-Th, Hep-Ph, *and* Epinions *overwhelmed IRM and MMSB. The ILA model did not finish even with our smallest dataset*

| | | **ML** | **AI/Robotics** | **Theory** |
|---|---|---|---|---|
| Nodes | | 2328 | 3417 | 1385 |
| Edges | | 3708 | 3788 | 1140 |
| Model | $K$ | | | |
| PCM | 10 | **0.71** ± 0.02 | **0.70** ± 0.02 | **0.75** ± 0.05 |
| IRM | (auto) | 0.59 ± 0.03 | 0.57 ± 0.03 | 0.59 ± 0.04 |
| MMSB | 10 | 0.50 ± 0.01 | 0.49 ± 0.03 | 0.50 ± 0.04 |
| SVI | 10 | 0.68 ± 0.01 | 0.67 ± 0.01 | 0.72 ± 0.04 |
| PCM | (opt) | **0.75** ± 0.02 | **0.73** ± 0.02 | **0.77** ± 0.05 |
| SVI | (opt) | 0.69 ± 0.01 | 0.67 ± 0.01 | 0.73 ± 0.04 |
| | | Hep-Th | Hep-Ph | Epinions |
| Nodes | | 27,770 | 34,546 | 75,879 |
| Edges | | 352,807 | 421,578 | 508,837 |
| Model | $K$ | | | |
| PCM | 10 | **0.85** ± 0.001 | 0.88 ± 0.001 | **0.86** ± 0.001 |
| IRM | (auto) | × | × | × |
| MMSB | 10 | × | × | × |
| SVI | 10 | × | **0.89** ± 0.001 | 0.81 ± 0.001 |
| PCM | (opt) | **0.93** ± 0.001 | **0.94** ± 0.001 | **0.88** ± 0.001 |
| SVI | (opt) | × | 0.93 ± 0.001 | 0.87 ± 0.001 |

of $A'(u)$. Ideally, the actual followers of $v$ in $E^{(\text{test})}$ would be ranked at the top of this list. We test this by computing the AUC score of the ranked list of predictions against the ground truth list which ranks the actual followers of $v$ in $E^{(\text{test})}$ at the top. This AUC score is averaged over all $v \in V^{(\text{test})}$ to yield a single score for PCM, with higher values indicating better link prediction accuracy, and hence better model fit. This test is then repeated with 30 train-test splits to get confidence bounds. This entire process is performed for all competing models, using the likelihood functions and parameters specific to those models.

Note that predicting the followers of a given producer (i.e., this test) is far more challenging than predicting the nodes followed by a given consumer, since the latter is strongly influenced by the *popularity* of producers (which can be easily inferred from their degree), while our test is not. Hence, it is a much more stringent test of the quality of topics found by the model.

We also note that while link prediction tasks are common for comparing network models, the goal of our paper is inferring the latent topics and not link prediction per se. Alternative methods of link prediction (e.g., by counting common neighbors between nodes, and variants [Adamic and Adar (2003), Katz (1953)] may be used when the node topics or clusters themselves are not desired.

*Accuracy.* We conducted two experiments. In the first experiment, we set all methods to use 10 topics/clusters (except for IRM, which automatically picks the number of topics); a limited number of clusters is important in situations where the topics or node interests must be visualized or processed further by an analyst. For PCM, we set the parameters as follows: (a) a hard constraint $\sum \theta_{ut} \leq 3$ of no more than 3 expected topics per node, and (b) a soft constraint of $\lambda = 0.5$ (for the sparser networks ML, AI/ROBOTICS, and THEORY) or $\lambda = 2.0$ (for the remaining denser networks). We note that these settings are *not optimized*; for example, the optimum settings for ML are $\lambda = 6$ with 25 topics, as shown later. This shows the outperformance of PCM for any reasonable parameters.

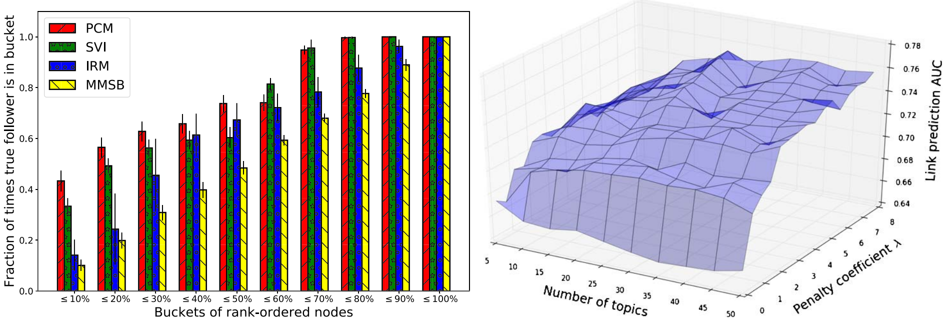
Table 1 shows the average AUC scores for all the models with 10 clusters. We observe that (a) PCM is better than all competing models for all datasets, except for HEP-PH where SVI is better, (b) the accuracy improvements are consistent over a wide variety of citation networks, and (c) PCM is able to *scale* to the larger datasets while several others either crashed or did not finish. Notice that the Mixed Membership Stochastic Blockmodel (MMSB) with MCMC inference performs quite poorly on these link prediction tasks.

In the second experiment, we take the two best methods from the previous experiment (PCM and SVI) and compare them on their optimal settings. These results mirror those of the previous experiment; PCM consistently does better.

Figure 2(a) shows a more detailed analysis: it counts the number of times the actual follower of a node $v \in V^{(\text{test})}$ was among the top-$k\%$ of the predicted followers of $v$, for the ML dataset. We see that (a) PCM has higher recall than the other models over a broad range, and (b) the greatest improvements are within the top-10% bucket, which consists of other nodes believed to have the same topic as $v$; this again points to the accuracy of the topics found by PCM.

*Parameter selection.* Figure 2(b) shows the effect of changing the number of topics and the penalty coefficient $\lambda$ on the accuracy of link prediction for ML. We see that accuracy is robust with respect to both parameters as long as $\lambda > 0$ (AUC is always between 0.7 and 0.75). For $\lambda = 0$, all consumers use the maximum allowed number of topics, which leads to a loss of accuracy. Hence, the soft constraint imposed by a positive $\lambda$ is necessary.

Another parameter of PCM is the maximum number of interests $\tau$ for any consumer (the hard constraint). Varying $\tau$ within a reasonable range of values tends to



(a) Recall (b) Effect of parameters on AUC

FIG. 2. (a) *For each test node $v$ in ML, all consumers are rank-ordered according to the probability of being a follower of $v$. The size of each bar represents the fraction of times the true followers of $v$ fall in the rank-ordered buckets. The true followers are in the top buckets significantly more often for PCM*. (b) *PCM is robust to parameter settings as long as* $\lambda > 0$ (*shown for ML*).

affect AUC scores by $\approx 0.01$, implying that a properly picked parameter for the soft constraint is good enough. The hard constraint can still be useful for robustness to outliers.

The above results show that the results of PCM are robust to the choice of parameters, as long as they are in the right range. In fact, link prediction accuracy could be used to choose the right parameter settings even for other applications of PCM. Indeed, this is precisely how we select the parameter $\lambda$ for our two applications, namely, inferring topics from Twitter, and from paper citations (Section 7). For these applications, we manually chose the number of topics $K$ for ease of presentation, but it could have been selected via link prediction as well.

*Scalability.* The updates for each node only require information from its neighbors, so the complexity is $O(K\Delta)$, where $K$ is the number of topics and $\Delta$ is the maximum degree of the network. In addition, PCM is easily parallelizable, by splitting node updates across multiple cores. The running times of the various models are shown in Table 2 (the parallelized version of PCM was used; the Mixed Mem-

TABLE 2
Running time: *We report total time for inference and predictions using* 10 *topics. Only PCM and SVI scale to the larger datasets*

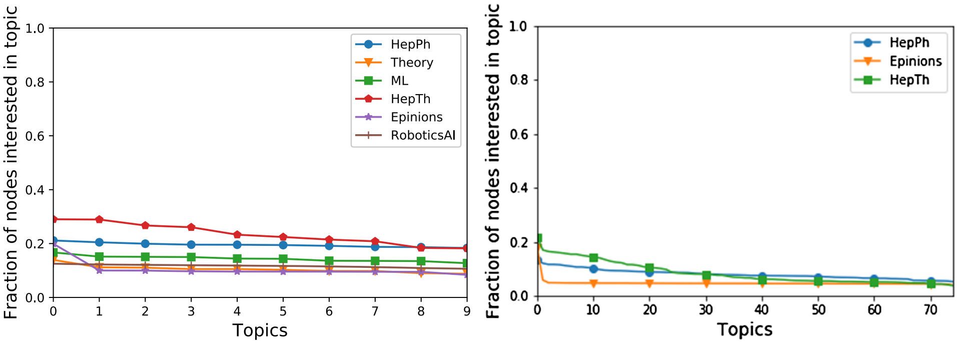
| | **ML** | **AI/ROBOTICS** | **THEORY** | **HEP-TH** | **HEP-PH** | **EPINIONS** |
|---|---|---|---|---|---|---|
| **PCM** | 26s | 32*s* | 13s | 1280s | 1565s | 2062s |
| **IRM** | 200s | 382s | 153s | × | × | × |
| **MMSB** | 2199s | 4637s | 739s | × | × | × |
| **SVI** | 5s | 5s | 3s | × | 2509s | 3240s |

bership Stochastic Blockmodel was run with 100 iterations and a burn-in of 100, as higher values increased run-time significantly). The scalability of PCM is clear.

*Topic size distribution.* Let the size of a topic denote the number of consumers interested in that topic. Formally, the size of a topic $t$ is $|\{u \in U \mid \theta_{ut} > 0.8\}|$, where we used a threshold value of 0.8 to indicate interest in consuming topic $t$. Figure 3 shows the distribution of topic size when a few (10) or many (75) topics are desired. In both cases, the ratio of sizes of the largest to the smallest topic is relatively small (within a factor of 2 for 10 topics, and a factor of 4 for 75 topics). This is because, for a "large" topic $t$, producers on that topic will typically have a small $\alpha_v$ since the set of followers of a producer of $t$ will not be "exclusive." More precisely, the followers will be drawn from a multinomial that is close to the uniform distribution, and hence yields a low likelihood. Thus, PCM implicitly penalizes large topics.

The behavior of PCM can be interpreted in terms of the trade-off between topic coherence and topic sizes. For instance, an algorithm that focuses on finding topics that are very well "separated" from each other may find a few such tightly-knit topics. However, a significant fraction of nodes may not belong to any of these topics. This results in at least one diffuse topic for such "left-over" nodes. Instead, PCM has a preference for similar-sized topics. Faced with a seemingly large-sized topic, PCM will try to find reasonably-sized sub-topics within it. Thus, each topic found by PCM is likely to be useful, with no diffuse topics for left-over nodes. The cost of this is that the topics found by PCM can be at different granularities, with especially popular topics being split up. We believe that this trade-off is worthwhile.

Note that we defined topic size in terms of the number of consumers interested in it, and not the number of producers. This is because we consider topic consumption to be a more natural measure of topic importance. The number of producers



(a) 10 topics (b) 75 topics

FIG. 3. *Fraction of consumers interested in each topic: A node u is interested in a topic if* $\theta_{ut} > 0.8$ (*thresholds between* 0.75 *and* 0.9 *were tested and give similar plots*). *Topic sizes are seen to be relatively similar.*

for the various topics are in fact more skewed, as we shall see in the Italian elections application in the next section.

## 7. Applications.

We now use PCM to analyze Twitter topics over the course of the 2013 Italian elections. Then we will find the main machine learning research topics from the citation network of papers published in machine learning conferences.

### 7.1. *Twitter topics during the Italian elections* (2013).

A general election took place on 24–25 February, 2013 to determine members of the Italian Chamber of Deputies and the Senate. While there were many parties involved, the primary ones were: (a) the *Con Monti per l'Italia (With Monti for Italy)* coalition, headed by the incumbent prime minister Mario Monti and his party (*Scelta Civica*), (b) the *PD* party, led by Pier Luigi Bersani, representing the center-left, (c) the *PdL* party, led by Silvio Berlusconi, representing the center-right, and (d) the *Five Star Movement (M5S)*, led by Beppe Grillo, a new entrant to the political scene, representing populist ideas.

*Data*: We use data collected by Caldarelli et al. (2014), available from http://www.linkalab.it/data. The data contains tweets from November 2012 until June 2013, but we ignore the first and last months (which have very low volume). All user names and tweets are anonymized, but the hashtags used in the tweets are available. We filter out users who received or sent fewer than 50 tweets in total. This yields, for each month, a directed graph of Twitter users ($A_{uv} = 1$ iff person $u$ tweeted "at" person $v$). Also, for each user, we know the hashtags used by that user in that month. Graph statistics are presented in Table 3.

*Experimental setup*: For a given month of data, we first apply PCM with $K = 5$ topics, and infer the latent topical interests of each user who tweeted in that month. We chose to set no hard threshold $\tau$ on the number of consumption interests, and the soft penalty parameter $\lambda$ was set to 2.0 by cross-validation. To interpret these topics, we find the best hashtags corresponding to these topics. This is *again* done via PCM. Specifically, we first build a directed network from users to hashtags (if $u$ used hashtag $h$, we create an edge $A_{uh} = 1$). Then, given the (known) topical interests of the users (the "consumers" in this network), we find the topical interest $t_h$ for each hashtag $h$ (the "producers"), and also its purity $\alpha_h$. Finally, for each

TABLE 3
*Graph statistics for Twitter data over the course of the Italian elections* (2013)

| | **December 2012** | **January 2013** | **February 2013** | **March 2013** | **April 2013** | **May 2013** |
|---|---|---|---|---|---|---|
| Nodes | 5061 | 6697 | 9360 | 5139 | 6197 | 2411 |
| Edges | 51,125 | 75,497 | 129,682 | 62,670 | 80,434 | 22,760 |

topic, we select the hashtags with the highest purity, that is, the hashtags that are most associated with the topic.[1]

*Description of results*: In the polarized climate engendered by elections, we expected topics to be aligned strictly along party lines and to be stable across time. However, neither of these turns out to be true. While some topics in each month may be party-specific, others can refer to coalitions (or even related parties). The topics that dominate the conversation change every month, as do the corresponding parties and politicians. Indeed, we find that the conversation on Twitter is heavily influences by TV talk shows and recent news/events. Mixed in with all of these are broader but ephemeral fads that arise occasionally on Twitter, and the influence of outside events on Italian politics (see Table 4).

*December 2012*: Three events dominate the discussion. Topic 1 follows Mr. Grillo (M5S) and his campaign for signatures (which he calls the "Massacre Tour") against Mr. Monti ("torna a casa, monti," or "back at home, Monti!"). Topic 3 follows Mr. Monti's efforts to engage with voters live on Twitter ("monti live") in favor of his party ("Scelta Civica"). Topic 4 is about the *Primarie delle idee* (Primaries of Ideas) event organized by the Fratelli d'Italia party in which they criticized Mr. Berlusconi's leadership of PdL; "senza paura" (i.e., "fearless") is the tag line. Topics 0 and 2 are more generic: the former aligns with Mr. Berlusconi ("I'm with Silvio") and the FLI party that split from Mr. Berlusconi's PdL, while the latter references the *Omnibusnotte*, an Italian TV show.

*January 2013*: A major topic of interest (Topic 1) is about a letter signed by Mr. Berlusconi delivered to Italian homes saying that he would abolish the IMU tax ("rimborso imu"). This topic is mixed with references to the "Lega Nord" (Northern League) regional political party whose leader was aligned with Mr. Berlusconi. Topic 4 combines information about the M5S party (the upcoming "Tsunami tour" of all parts of Italy by Mr. Grillo) and the Fratelli d'Italia party (and its cofounder Ms. Meloni). Topics 2 and 3 references the *Le Invasioni Barbariche* TV talk show.

*February 2013*: In the lead-up to the elections, held at the end of February, we find an increasing interest in the niche party called *FARE*, which is represented alongside leaders Mr. Boldrin and Ms. Silvia Enrico in topics 2 and 4. Topic 0 refers to the *Scelta Civica* party of Mr. Monti, and the coalition of allied parties (*Con Monti per l'Italia*). We also see the growth of a Twitter fad, namely, the *indivanados* hashtag, which originated as a tweet-based chain-letter. Finally, again we see a TV program (*Ultima Parola*) significantly affecting the topics of interest of Twitter users.

*March 2013*: The elections failed to deliver a decisive result. The PD party, led by Mr. Bersani, was asked to form a coalition, but failed to do so. The Twitter discussion is dominated by small parties which could affect coalitions, and by

[1] We only consider the top-200 hashtags by frequency to avoid selecting esoteric hashtags. Topic descriptions are based on detailed searches of online media based on the hashtags.

TABLE 4
*Twitter topics over the course of the Italian elections* (2013). *For each topic*, *the primary hashtags and their counts are shown*

| | |
|---|---|
| | December 2012 |
| Topic 0 | #iostoconsilvio (208) #montibis (211) #berlusconi2013 (143) #ue (171) #fli (205) |
| Topic 1 | nohashtag (272) #massacrotour (182) #firmaday (298) #crisigoverno (150) #tornaacasamonti (423) |
| Topic 2 | #scalfari (147) #omnibusnotte (157) #report (305) #sischerza (192) #montibis (211) |
| Topic 3 | #tvb (150) #fiat (146) #primarieparlamentari (425) #sceltacivica (663) #montilive (1001) |
| Topic 4 | #primariedelleidee (162) #primarieidee (224) #senzapaura (371) #centrodestra (162) #fratelliditalia (407) |
| | January 2013 |
| Topic 0 | nohashtag (414) #agendatweet (208) #iostoconambrosoli (253) #shoah (324) #lavoro (676) |
| Topic 1 | #rimborsoimu (233) #leganord (221) #cgil (286) #ppe (203) #iostoconsilvio (256) |
| Topic 2 | #seviziapubblica (452) #leinvasioni (241) #invasioni (576) #rassegnati (211) #firenze (331) |
| Topic 3 | #carfagna (189) #sischerza (186) #leinvasionibarbariche (215) #coerenza (193) #ff (223) |
| Topic 4 | #rai3 (224) #iovotom5s (306) #fratelliditalia (634) #tsunamitour (1357) #meloni (386) |
| | February 2013 |
| Topic 0 | #conmontiperlitalia (384) #sceltacivica (672) #letta (289) #ariachetira (422) #faresulserio (534) |
| Topic 1 | nohashtag (501) #instantpoll (418) #padova (264) #conmontiperlitalia (384) #tg1 (266) |
| Topic 2 | #boldrin (303) #combattere (278) #indivanados (330) #grecia (304) #silviaenrico (370) |
| Topic 3 | #fratelliditalia (575) #iovotom5s (462) #udc (353) #dipietro (512) #sel (875) |
| Topic 4 | #ultimaparola (559) #indivanados (330) #boldrin (303) #sallusti (542) #iovotofare (337) |
| | March 2013 |
| Topic 0 | nohashtag (384) #nuovogoverno (126) #elezioni (192) #nuovecamere (133) #aldrovandi (118) |
| Topic 1 | #pude (106) #sceltacivica (187) #ff (156) #fornero (127) #sischerza (182) |
| Topic 2 | #m5s! (104) #8marzo (155) #prodi (148) #poveropaese (143) #bersani! (124) |
| Topic 3 | #rivoluzionecivile (112) #ghedini (115) #campanella (142) #8punti (108) #fumatabianca (167) |
| Topic 4 | #openpd (122) #rassegnati (188) #pude (106) #trasparenza (125) #lega (290) |
| | April 2013 |
| Topic 0 | #adesso (162) #italy (135) #presidenteditutti (331) #fratelliditalia (151) #chiamparino (171) |
| Topic 1 | #chetempochefa (187) #25aprile (148) #ballaro (307) #info5stelle (146) #boston (140) |
| Topic 2 | #iostoconbersani (177) #perdire (437) #b (175) #raisenzapartiti (257) #controlapovertà (280) |
| Topic 3 | #casta (207) #crisi (193) #elezioni (170) #sicilia (142) #noprodi (303) |
| Topic 4 | nohashtag (367) #rodotàperchèno (205) #fiatosulcolle (267) #politica (505) #rodotàperchéno (454) |

TABLE 4
(*Continued*)

| | May 2013 |
|---|---|
| Topic 0 | #carfagna (105) #brunetta (129) #santanchè (129) #salto13 (118) #ineleggibilità (104) |
| Topic 1 | #tuttiacasa (125) #eleroma (160) #nonsiamounpartito (336) #leggetruffa (221) #tuttiacasatour (519) |
| Topic 2 | nohashtag (193) #ultimora (134) #giustizia (148) #lega (115) #roma (298) |
| Topic 3 | #openpd (144) #oltrelarottamazione (245) #salto13 (118) #firenze (130) #assembleapd (425) |
| Topic 4 | #ingroia (289) #grillo? (166) #falcone (111) #italia (113) #finocchiaro (125) |

internal strife within the PD party. For instance, Topics 1 and 3 reference the *Scelta Civica* party of Mr. Monti, and the small *Rivoluzione Civile* party, respectively. However, we also find the hashtag `#ff` (or *Follow Friday*), a long-standing Twitter tradition of users pointing out interesting new accounts or hashtags to their friends. Topic 4 is primarily about strife within the PD party and the call for an *open PD* with new leadership and calls for resignation (*rassegnati*).

*April 2013*: With the governmental succession in chaos, a new election to choose the president was called in April. The president would be chosen by delegates, not by the public directly. After the PD's candidate failed to win after several ballots, PD split and the incumbent leader Mr. Bersani resigned. Finally, the incumbent president was re-elected, and he encouraged PD's deputy secretary Mr. Letta to form a coalition, which he duly accomplished by the end of April.

The topics for April reflect the sense of crisis in the elections (Topic 3), and the interest in selecting someone who could be a "president of all" (*presidenteditutti*; Topic 0). TV talk shows again take center stage (*Che Tempo Che Fa* and *Ballaro* in Topic 1). Recent news events also get significant attention (Topic 2), as evidenced by interest in Mr. Bersani's rally on April 13 "against poverty" (*contro la poverta*) and Mr. Grillo's proposal to "free TV channel RAI from the parties" (*RAI senza partiti*).

*May 2013*: The three main parties (PD, PdL, and M5S) were all affected differently by the election results. Topic 0 documents turmoil regarding leadership of the PdL after it appeared that the incumbent leader (Mr. Berlusconi) may be convicted. There was heated discussion regarding the possible candidacies of Ms. Carfagna, Ms. Santanche, and Mr. Brunetta as possible replacements. Topic 1 refers to the leader of M5S (Mr. Grillo) encouraging an effort against other parties under the tagline *tutti a casa* ("send them all home"). It also notes his efforts against a proposed law sponsored by PD that would exclude organizations (such as M5S) that were not technically parties from receiving public funding. He railed against the proposal saying that M5S was not a party (*non siamo un partito*) and would never become one. The PD also faced calls for new leadership (Topic 3), expressed by the hashtag `#openpd`, after the resignation of Mr. Bersani. A leading candidate (Mr.

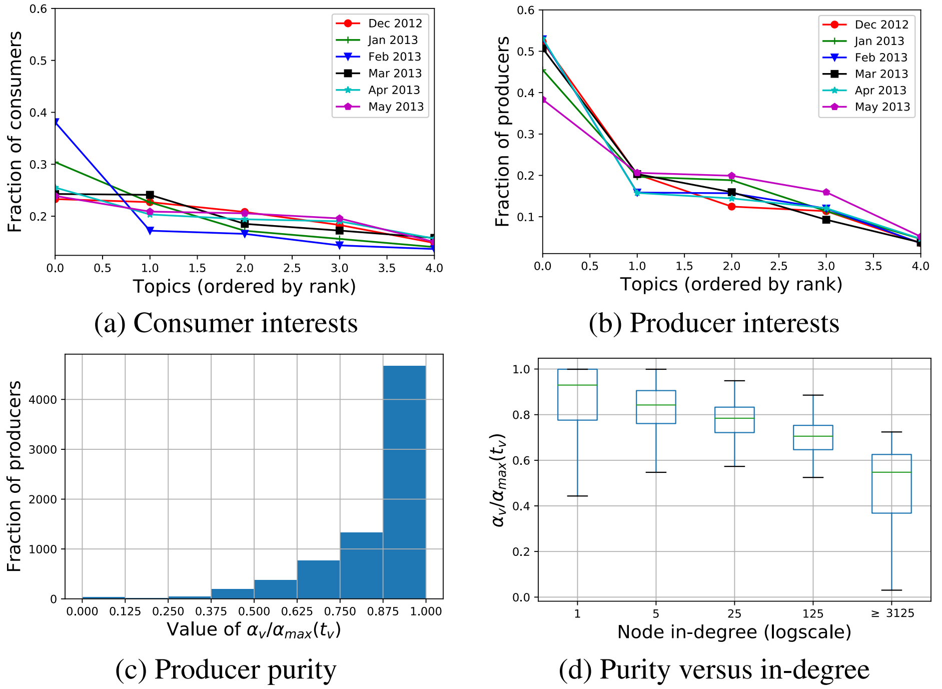
Renzi) publishing a book, called *Oltre la Rottamazione* ("beyond the scrapping") in May 19, 2013, which gained much attention.

*Summary of Twitter topics*: Thus, we find that in contrast to our expectations, the topics are neither stable across time, nor are they always aligned along party lines. The topics fluctuate depending on events promoted by parties (e.g., the *tsunami tour* of M5S, or the *contro la poverta* rally of PD), and also the talking points discussed on Italian TV shows (e.g., *Le Invasioni Barbariche*). Some topics possess a degree of continuity, but interest in them can vary greatly over time. For instance, the `#openpd` hashtag was popular in March 2013 and also in May 2013, but not in the intervening month. Also, the most discriminative hashtags are also not necessarily the most common ones. For instance, `#berlusconi` is used by users across the political spectrum. Picking hashtags $h$ with high purity scores $\alpha_h$ allows PCM to solve this problem, and create interpretable topic summaries.

*Analysis of inferred node parameters*: Figure 4 shows several facets about the inferred parameters. The number of consumers and producers interested in the various topics are shown in plots (a) and (b). In each, the topics are ordered according to total consumer (producer) interest. Recall from Section 6 (Figure 3, and the discussion on "Topic size distribution") that PCM tends to find topics with balanced consumption interests. This is because any topic that is of interest to a large fraction of consumers will have an associated multinomial distribution that is close to uniform, which in turn leads to a lower likelihood. However, there is no such penalty for the number of producers tweeting about any topic. This is validated by our results, where consumer interest varies only by a factor of 2 across topics [plot (a)], but shows greater variation for producer interest [plot (b)].

In Figure 4(c), we show the distribution node purity values $\alpha_v$, normalized by the maximum possible purity for those topics. We find that most producers are extremely pure, that is, most of their followers have a consumption interest in the topic of their tweets. In plot (d), we split the producers by their in-degrees, and compute purity for various in-degree ranges. This shows a slightly downward trend in purity as in-degree increases, suggesting that as people become more popular, they can gain followers for reasons unrelated to the topic of their tweets (e.g., celebrities may gain followers simply because they are famous). While these two plots are for only February 2013, we find that every month shows the same pattern.

Finally, we also computed the number of consumption interests for each Twitter user. As in the computation of topic sizes, we say that a user is interested in consuming topic $t$ if $\theta_{ut} > 0.8$. We find that almost half the users are interested in only one topic, and about 65% in up to two topics. This shows that a significant fraction of users indeed have few consumption interests. This fact, coupled with the observation of high average purity discussed earlier, suggests that PCM is operating in the regime where it performed particularly well in the simulation experiments.



(a) Consumer interests

(b) Producer interests

(c) Producer purity

(d) Purity versus in-degree

FIG. 4. (a) *Topics are ordered in descending order of the number of consumers interested in them* (*a consumer* $u$ *is interested in topic* $t$ *if* $\theta_{ut} > 0.8$). *We find little difference between topic sizes*, *agreeing with Figure* 3. (b) *On ordering topics by the number of producers tweeting about those topics*, *we find a greater range of topic sizes*. (c) *The distribution of node purity* $\alpha_v$ *for the election month* (*February* 2013) *shows that most have high purity* [*i.e.*, $\alpha_v \approx \alpha_{\max}(t_v)$, *where* $\alpha_{\max}(t)$ *is the maximum allowable purity for a topic* $t$]. *Thus*, *most producers can indeed be modeled using a single topic*. (d) *Plotting* $\alpha_v$ *versus in-degree* (*for February* 2013), *we find a slightly downward trend*. *People with too many followers are less pure*, *that is*, *their followers are not all interested in the topic of their tweets* (*see also Remark* 1).

7.2. *Research areas extracted from a citation network.* We use PCM to extract 10 topics from the ML citation network. Then, as with the Twitter topics above, we find the best words for each topic by (a) building a directed network from papers to words and phrases extracted from paper titles, (b) calculating the "production" interest of each word from the known "consumption" interests of all papers pointing to it, and finally (c) selecting the words with the highest purity for each topic as the representatives for that topic.

Table 5 lists these representative keywords for each topic (only words that occur at least 30 times are considered). Two topics are devoted to reinforcement learning, demonstrating its importance to planning and AI. Topic 1 alludes to the policy gradient and function approximation methods for reinforcement learning, while Topic 8 is about kernel-based methods. State-space representations and Markov Decision Processes also garner two topics (Topics 2 and 7). Other topics are asso-

TABLE 5

*Example topics for ML: PCM was used to find 10 topics explaining the machine-learning paper citation network. For each topic, the best matching words extracted from paper titles were found, again via PCM. Both the top 5 words and the corresponding $\alpha_v$ values are reported (higher $\alpha_v$ implies stronger association of the word with the topic)*

| | |
|---|---|
| Topic 0 (336 papers) | Gaussian_process (2.94) active (1.71)<br>support_vector_machines (1.56) sparse (1.52) kernel (1.22) |
| Topic 1 (306 papers) | policy (1.95) gradient (1.72)<br>approximation (1.41) fast (1.29) function (0.98) |
| Topic 2 (275 papers) | state (1.36) representations (1.21)<br>boosting (0.86) process (0.81) bounds (0.73) |
| Topic 3 (279 papers) | semi-supervised (2.18) kernels (2.11)<br>information (1.50) clustering (1.09) probabilistic (0.94) |
| Topic 4 (348 papers) | policy (2.00) approximation (1.87)<br>matrix (1.75) function (1.36) dynamic (1.15) |
| Topic 5 (303 papers) | local (1.00) search (1.00)<br>multiple (0.99) recognition (0.97) optimal (0.96) |
| Topic 6 (323 papers) | modeling (0.87) sparse (0.66)<br>boosting (0.62) machine (0.59) models (0.50) |
| Topic 7 (327 papers) | representations (1.80) semi-supervised (1.70)<br>state (1.31) stochastic (1.09) clustering (0.95) |
| Topic 8 (321 papers) | kernels (1.66) reinforcement_learning (1.27)<br>approximation (1.18) functions (0.79) markov (0.58) |
| Topic 9 (319 papers) | matrix (1.17) semi-supervised (1.17)<br>prediction (1.07) stochastic (0.93) kernel (0.90) |

ciated with disparate research areas such as Gaussian processes (Topic 0), semi-supervised methods (Topic 3), local search methods (Topic 5), and matrix-based methods (Topic 9). Topic 4 is a combination of function approximation (Topic 1) and matrix methods (Topic 9). The only generic topic is Topic 6, tagged by common keywords such as `modeling` and `sparse`. Note that the purity values are relatively low for Topic 6, and this can be used to automatically infer that this topic is less coherent, and hence less important. All the other topics are coherent, and have at least some words with high purity values.

Table 5 also shows the total consumer interest in each topic, in terms of the number of papers which have $\theta_{ut} > 0.8$ in that particular topic. We find that the sizes of all topics are fairly similar, echoing results from Figures 3 and 4. As noted earlier, this reflects the tendency of PCM to find topics at different granularities, such that each has similar levels of consumer interest.

## 8. Conclusions.

Networks resulting from human actions reflect the incentives of the individuals involved, and directed networks may differ from their undirected counterparts in this regard. In particular, the production and consumption interests of nodes may differ, and each producer may be known to consumers as an expert on only one topic. We presented the PCM model for directed networks that reflects such incentives. We developed a fast alternating-optimization procedure for parameter inference under PCM. Experiments on simulated data as well as several real-world datasets showed that PCM significantly outperforms existing models both in terms of accuracy as well as scalability.

We then used PCM to explore the topics of interest to Twitter users during the Italian elections of 2013. In addition, we tagged these topics with the most relevant hashtags, again using PCM. The results show that Twitter discussions are not necessarily aligned along party lines, but rather focus on recent political events. The reach and importance of TV talk shows is also clear. As a second application, we used PCM to infer topics of research from a machine learning paper citation network, and tag them with appropriate keywords extracted from paper titles. This again finds intuitive and interpretable topics.

One potential application of PCM is in search and recommendation systems. Consider a new Twitter user who searches for a particular topic. Twitter should ideally return links to other Twitter users who are "authorities" on that topic. However, an authority is not necessarily one with many followers. Indeed, as we have seen, those with very high in-degrees often have low node purity, that is, their followers are not specifically interested in the topics of their tweets, but are perhaps only interested in following celebrities. An authority must ideally have high purity as well as high popularity, signifying that the person is widely followed (and hence endorsed) primarily by those those interested in a topic. In fact, we used this very principle when picking hashtags to represent topics: we selected from among the most popular hashtags the ones that had the highest purity for each topic. Extending and applying this to a general recommendation system is an interesting direction for future work.

One limitation of PCM is our inability to estimate parameter distributions; our analysis only yields point estimates. The standard solution via a bootstrap is difficult here, because the network is not a collection of i.i.d. samples of nodes and links. The "network bootstrap" is an active area of research, but we are unaware of any method with provable guarantees. We note that this problem is common to all network models and not specific to PCM. Even for the well-studied Mixed-Membership Stochastic Blockmodel, only the MCMC-based methods yield distributions. However, as we showed in Section 6 (Table 2), this is difficult to scale to large networks.

Another possible concern about PCM is that there may be Twitter users who choose to tweet about multiple topics in spite of their incentives to focus on just one topic. First, we note that the PCM generative model of network structure remains valid even in this case as long as users are "known" for only one topic, that is, they attract followers for their tweets on only that topic. If even this weaker assumption fails to hold, PCM will pick up the primary topic of producer $v$ as her single production interest. The fact that $v$ has followers who are interested in other topics will be reflected in a lower node purity $\alpha_v$. In the extreme case where a producer tweets about a wide range of topics, we will have $\hat{\alpha}_v \approx 0$, and the contribution of $v$ and her followers to the likelihood will become nearly constant. Thus, a few such atypical Twitter users will not affect inference. Indeed, inference of consumer interests will be mainly driven by "pure" nodes who have a single interest. The importance of pure nodes is also emphasized in recent work on a variant of the Mixed Membership Stochastic Blockmodel [Zhang, Levina and Zhu (2014)].

Also, while we have proven dyadic effects such as reciprocity, network models often have higher-order effects as well. Proving these effects for PCM is difficult. Some latent space models, such as the one by Hoff, Raftery and Handcock (2002), automatically yield networks with transitivity. However, they are not as scalable as PCM. Hence, we believe that PCM is primarily applicable to large social network datasets.

Finally, we have not proven the consistency of the MAP estimator for PCM. The difficulty is that each node has parameters, and hence the parameter size grows with the number of nodes. This is in general a difficult problem, and recent work has shown consistency for random graphs with known degrees [Chatterjee, Diaconis and Sly (2011)], for directed exponential random graphs with known in- and out-degrees [Yan, Leng and Zhu (2016)], and for exponential random graphs under sampling with certain conditions [Shalizi and Rinaldo (2013)]. For stochastic blockmodels and variants, consistency has been proven for alternative inference methods, for example, spectral clustering for stochastic blockmodels [Lei and Rinaldo (2015)], and the OCCAM method for a variant of the Mixed Membership Stochastic Blockmodel [Zhang, Levina and Zhu (2014)]. However, for PCM, this is an area of future work.

## SUPPLEMENTARY MATERIAL

**Supplement A: Proofs** (DOI: 10.1214/17-AOAS1079SUPP; .pdf). We provide the proofs for all propositions and theorems.

## REFERENCES

ADAMIC, L. and ADAR, E. (2003). Friends and neighbors on the Web. *Soc. Netw.* **25** 211–230.

AIELLO, W., CHUNG, F. and LU, L. (2000). A random graph model for massive graphs. In *Proceedings of the Thirty-Second Annual ACM Symposium on Theory of Computing* 171–180. ACM, New York. MR2114530

AIROLDI, E. M., BLEI, D. M., FIENBERG, S. E. and XING, E. P. (2008). Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.* **9** 1981–2014.

BLEI, D. M., NG, A. Y. and JORDAN, M. I. (2003). Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3** 993–1022.

CAIMO, A. and FRIEL, N. (2011). Bayesian inference for exponential random graph models. *Soc. Netw.* **33** 41–55. DOI:10.1016/j.socnet.2010.09.004.

CALDARELLI, G., CHESSA, A., PAMMOLLI, F., POMPA, G., PULIGA, M., RICCABONI, M. and RIOTTA, G. (2014). A multi-level geographical study of Italian political elections from Twitter data. *PLoS ONE* **9** e95809.

CARAGEA, C., WU, J., CIOBANU, A., WILLIAMS, K., FERNÁNDEZ-RAMÍREZ, J., CHEN, H.-H., WU, Z. and GILES, L. (2014). CiteSeerX: A scholarly big dataset. In *Proceedings of the* 36*th European Conference on Information Retrieval* (*ECIR'*14) 311–322.

CHAKRABARTI, D. (2017). Supplement to "Modeling node incentives in directed networks." DOI:10.1214/17-AOAS1079SUPP.

CHAKRABARTI, D. and FALOUTSOS, C. (2006). Graph mining: Laws, generators, and algorithms. *ACM Comput. Surv.* **38** Article No. 2. DOI:10.1145/1132952.1132954.

CHAKRABARTI, D., ZHAN, Y. and FALOUTSOS, C. (2004). R-MAT: A recursive model for graph mining. In *Proceedings of the* 4*th SIAM International Conference on Data Mining* (*SDM'*04) 442–446.

CHANG, J. (2012). lda: Collapsed Gibbs sampling methods for topic models. Available at https://cran.r-project.org/web/packages/lda/index.html.

CHATTERJEE, S., DIACONIS, P. and SLY, A. (2011). Random graphs with a given degree sequence. *Ann. Appl. Probab.* **21** 1400–1435. DOI:10.1214/10-AAP728.

DUIJN, M. A., SNIJDERS, T. A. and ZIJLSTRA, B. J. (2004). P2: A random effects model with covariates for directed graphs. *Stat. Neerl.* **58** 234–254.

ERDŐS, P. and RÉNYI, A. (1959). On random graphs. I. *Publ. Math. Debrecen* **6** 290–297. MR0120167

FOSDICK, B. K. and HOFF, P. D. (2015). Testing and modeling dependencies between a network and nodal attributes. *J. Amer. Statist. Assoc.* **110** 1047–1056. MR3420683

FRANK, O. and STRAUSS, D. (1986). Markov graphs. *J. Amer. Statist. Assoc.* **81** 832–842. MR0860518

FU, W., SONG, L. and XING, E. P. (2009). Dynamic mixed membership blockmodel for evolving networks. In *Proceedings of the* 26*th Annual International Conference on Machine Learning* (*ICML'*09) 329–336.

GEHRKE, J., GINSPARG, P. and KLEINBERG, J. (2003). Overview of the 2003 KDD cup. *ACM SIGKDD Explor. Newsl.* **5** 149–151. DOI:10.1145/980972.980992.

GILBERT, E. N. (1959). Random graphs. *Ann. Math. Stat.* **30** 1141–1144. MR0108839

GOPALAN, P. K. and BLEI, D. M. (2013). Efficient discovery of overlapping communities in massive networks. *Proc. Natl. Acad. Sci. USA* **110** 14534–14539. MR3105375

HANDCOCK, M. S. and JONES, J. H. (2004). Likelihood-based inference for stochastic models of sexual network formation. *Theor. Popul. Biol.* **65** 413–422. DOI:10.1016/j.tpb.2003.09.006.

HOFF, P. D. (2005). Bilinear mixed-effects models for dyadic data. *J. Amer. Statist. Assoc.* **100** 286–295. MR2156838

HOFF, P. D. (2009). Multiplicative latent factor models for description and prediction of social networks. *Comput. Math. Organ. Theory* **15** 261–272.

HOFF, P. D., RAFTERY, A. E. and HANDCOCK, M. S. (2002). Latent space approaches to social network analysis. *J. Amer. Statist. Assoc.* **97** 1090–1098. MR1951262

HOFMANN, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the* 22*nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR'*99) 50–57.

HOFMANN, T. (2004). Latent semantic models for collaborative filtering. *ACM Trans. Inf. Syst.* **22** 89–115. DOI:10.1145/963770.963774.

HOLLAND, P. W., LASKEY, K. B. and LEINHARDT, S. (1983). Stochastic blockmodels: First steps. *Soc. Netw.* **5** 109–137. MR0718088

HOLLAND, P. W. and LEINHARDT, S. (1981). An exponential family of probability distributions for directed graphs. *J. Amer. Statist. Assoc.* **76** 33–65. MR0608176

HUNTER, D. R. and HANDCOCK, M. S. (2006). Inference in curved exponential family models for networks. *J. Comput. Graph. Statist.* **15** 565–583. MR2291264

KARRER, B. and NEWMAN, M. E. J. (2011). Stochastic blockmodels and community structure in networks. *Phys. Rev. E* (3) **83** 016107. MR2788206

KATZ, L. (1953). A new status index derived from sociometric analysis. *Psychometrika* **18** 39–43.

KEMP, C., TENENBAUM, J. B., GRIFFITHS, T. L., YAMADA, T. and UEDA, N. (2006). Learning systems of concepts with an infinite relational model. In *Proceedings of the* 21*st National Conference on Artificial Intelligence* (*AAAI*'06) 381–388.

KOREN, Y. (2008). Factorization meets the neighborhood: A multifaceted collaborative filtering model. In *Proceedings of the* 14*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (*KDD*'08) 426–434.

KRIVITSKY, P. N., HANDCOCK, M. S., RAFTERY, A. E. and HOFF, P. D. (2009). Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Soc. Netw.* **31** 204–213. DOI:10.1016/j.socnet.2009.04.001.

LEI, J. and RINALDO, A. (2015). Consistency of spectral clustering in stochastic block models. *Ann. Statist.* **43** 215–237. MR3285605

LESKOVEC, J., CHAKRABARTI, D., KLEINBERG, J., FALOUTSOS, C. and GHAHRAMANI, Z. (2010). Kronecker graphs: An approach to modeling networks. *J. Mach. Learn. Res.* **11** 985–1042. MR2600637

MILLER, K. T., GRIFFITHS, T. L. and JORDAN, M. I. (2009). Nonparametric latent feature models for link prediction. In *Advances in Neural Information Processing Systems* 22 (*NIPS*'09) 1276–1284.

PALLA, K., KNOWLES, D. A. and GHAHRAMANI, Z. (2012). An infinite latent attribute model for network data. In *Proceedings of the* 29*th International Conference on Machine Learning* (*ICML*'12) 1607–1614.

RAFTERY, A. E., NIU, X., HOFF, P. D. and YEUNG, K. Y. (2012). Fast inference for the latent space network model using a case-control approximate likelihood. *J. Comput. Graph. Statist.* **21** 901–919. MR3005803

RICHARDSON, M., AGRAWAL, R. and DOMINGOS, P. M. (2003). Trust management for the semantic web. In *Proceedings of the* 2*nd International Semantic Web Conference* (*ISWC*'03) 351–368.

SALTER-TOWNSHEND, M. and MURPHY, T. B. (2013). Variational Bayesian inference for the latent position cluster model for network data. *Comput. Statist. Data Anal.* **57** 661–671. MR2981116

SARKAR, P. and MOORE, A. W. (2010). Fast nearest-neighbor search in disk-resident graphs. In *Proceedings of the* 16*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (*KDD*'10) 513–522.

SHALIZI, C. R. and RINALDO, A. (2013). Consistency under sampling of exponential random graph models. *Ann. Statist.* **41** 508–535. MR3099112

SNIJDERS, T. A. B. and NOWICKI, K. (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *J. Classification* **14** 75–100. MR1449742

VU, D. Q., HUNTER, D. R. and SCHWEINBERGER, M. (2013). Model-based clustering of large networks. *Ann. Appl. Stat.* **7** 1010–1039. MR3113499

WANG, Y. J. and WONG, G. Y. (1987). Stochastic blockmodels for directed graphs. *J. Amer. Statist. Assoc.* **82** 8–19. MR0883333

WASSERMAN, S. and PATTISON, P. (1996). Logit models and logistic regressions for social networks. I. An introduction to Markov graphs and $p$. *Psychometrika* **61** 401–425. MR1424909

XU, Z., TRESP, V., YU, K. and KRIEGEL, H. (2006). Infinite hidden relational models. In *Proceedings of the* 22*nd Conference on Uncertainty in Artificial Intelligence* (*UAI*'06) 544–551.

YAN, T., LENG, C. and ZHU, J. (2016). Asymptotics in directed exponential random graph models with an increasing bi-degree sequence. *Ann. Statist.* **44** 31–57. MR3449761

ZHANG, Y., LEVINA, E. and ZHU, J. (2014). Detecting overlapping communities in networks using spectral methods. ArXiv e-print. Available at https://arxiv.org/abs/1412.3432.

MCCOMBS SCHOOL OF BUSINESS
INFORMATION, RISK, AND OPERATIONS MANAGEMENT (IROM)
UNIVERSITY OF TEXAS, AUSTIN
CBA 6.462
2110 SPEEDWAY
STOP B6500
AUSTIN, TEXAS 78712
USA
E-MAIL: deepay@utexas.edu