Exploratory Study of a New Model for Evolving Networks

Anna Goldenberg and Alice Zheng

Carnegie Mellon University, Pittsburgh, PA 15213, USA
anya@cs.cmu.edu,alicez@cs.cmu.edu

Abstract. The study of social networks has gained new importance with the recent rise of large on-line communities. Most current approaches focus on deterministic (descriptive) models and are usually restricted to a preset number of people. Moreover, the dynamic aspect is often treated as an addendum to the static model. Taking inspiration from real-life friendship formation patterns, we propose a new generative model of evolving social networks that allows for birth and death of social links and addition of new people. Each person has a distribution over social interaction spheres, which we term "contexts." We study the robustness of our model by examining statistical properties of simulated networks relative to well known properties of real social networks. We discuss the shortcomings of this model and problems that arise during learning. Several extensions are proposed.

1 Introduction

In 1967, the seminal "small world" study [1] brought social networks into the public consciousness. Since then, researchers have paid close attention to laws that seem to govern human and business networks. How do links between people form? Is it enough to look at pairs or should triads of individuals be considered separately? Many approaches study networks on the scale of links and individuals to identify key patterns and describe network properties [2].

Data collection used to be an expensive and tedious process prone to sampling bias. But as more information are becoming available on-line, networks on the order of tens of thousands of people have become easily accessible. Studies of large hyper-link networks reveal similar behavior to those of large social nets (e.g. co-authorships). Thus a new modeling approach was developed based on random graphs [3, 4]. Here the goal is not to model the network on a link-by-link basis but to address its overall behavior. The new approach is more generative in nature, though most models are still very simplistic. The preferential attachment model [3] describes the mechanism of network evolution with a focus on power-law degree distributions. Once the links are established, they remain in the network unperturbed. Such simplifying assumptions make the models feasible for analysis, but fail to capture the complexity of real social networks.

In this work, we attempt to address several important issues raised by both communities. First, we directly model the generative process behind network dynamics. We focus on the evolution of interpersonal relationships over time, and explicitly model the birth and gradual decay of social links. Secondly, we demonstrate that the model generates networks that exhibit properties commonly observed in many natural topologies.

We motivate our model with an example. Imagine that Andy moves to a new town. He may find some new collaborators at work, make friends at parties, or meet fellow gym-goers while exercising. In general, Andy lives in a number of different spheres of interaction or *contexts*. As time goes on, he may find himself repeatedly meeting certain people in different contexts, consequently developing stronger bonds. Acquaintances he never meets again may quickly fade away. Andy's new friends may also introduce him to their friends (a well known transitive phenomenon called *triadic closures* in social science [2]).

With this example in mind, we begin with a presentation of our model in Section 2. Experimental results are discussed in Section 3. We show how to learn the parameters of our model using Gibbs sampling in Section 4, and give possible extensions of the model in Section 5. Section 6 contains a brief survey of related work, and Section 7 discusses the strengths and weaknesses of the proposed model.

2 The Model

2.1 Notation

DCFM allows the addition of new people into the network at each time step. Let T denote the total number of time steps and N_t the number of people at time t. $N = N_T$ denotes the final total number of people. Let M_t denote the number of new people added to the network at time t, so that $N_t = N_{t-1} + M_t$.

Links between people are weighted. Let $\{W^1, \ldots, W^T\}$ be a sequence of weight matrices, where $W^t \in \mathbb{Z}_+^{N_t \times N_t}$ represents the pairwise link weights at time t. We assume that W^t is symmetric, though it can be easily generalized to the directed case.

The intuition behind our model is that friendships are formed in *contexts*. There are a fixed number of contexts in the world, K, such as work, gym, restaurant, grocery store, etc. Each person has a distribution over these contexts, which can be interpreted as the average percentage of time that he spends in each context.

2.2 The Generative Process

At time t, the N_t people in the network each selects his current context R_i^t from a multinomial distribution with parameter θ_i, where θ_i has a Dirichlet prior distribution:

$$\boldsymbol{\theta}_i \sim \mathrm{Dir}(\boldsymbol{\alpha}), \quad \forall i = 1 : N \tag{1}$$

$$R_i^t \mid \theta_i \sim \mathrm{Mult}(\theta_i), \quad \forall t = 1 : T, i = 1 : N_t. \tag{2}$$

The number of all possible pairwise meetings at time t is $\mathrm{DYAD}^t = \{(i,j) \mid 1 \leq i \leq N_t, i < j \leq N_t\}$. For each pair of people i and j who are in the same context at time t (i.e., $R_i^t = R_j^t$), we sample a Bernoulli random variable F_{ij}^t with parameter $\beta_i\beta_j$. If $F_{ij}^t = 1$, then i and j meets at time t. The parameter β_i may be interpreted as a measurement of friendliness and is a beta-distributed random variable (making it possible for people to have different levels of friendliness):

$$\beta_i \sim \mathrm{Beta}(a,b), \quad \forall i = 1:N, \quad \forall (i,j) \in \mathrm{DYAD}^t$$

$$F_{ij}^t \mid R_i^t, R_j^t \sim \begin{cases} \mathrm{Ber}(\beta_i\beta_j) & \text{if } R_i^t = R_j^t \\ I_0 & \text{o.w.} \end{cases} \tag{3}$$

where I_0 is the indicator function for $F_{ij}^t = 0$.

In addition, the newcomers at time t have the opportunity to form triadic closures with existing people. The probability that a newcomer j is introduced to existing person i is proportional to the weight of the links between i and the people whom j meets in his context. Let $\mathrm{TRIAD}^t = \{(i,j) \mid 1 \leq i \leq N_{t-1}, 1 \leq j \leq M_t\}$ denote the pairs of possible triadic closures. For all $(i,j) \in \mathrm{TRIAD}^t$, we have:

$$G_{ij}^t \mid W^{t-1}, F_{\cdot j}^t, R_{\cdot}^t \sim \begin{cases} \mathrm{Ber}(\mu_{ij}^t) & \text{if } R_i \neq R_j \\ I_0 & \text{o.w.}, \end{cases} \tag{4}$$

where $\mu_{ij}^t := \sum_{\ell=1}^{N_t} W_{i\ell}^{t-1} F_{\ell j}^t / \sum_{\ell=1}^{t-1} W_{i\ell}^{t-1}$.

Connection weight updates are Poisson distributed. Our choice of a discrete distribution allows for sparse weight matrices, which are often observed in the real world. Pairwise connection weights may drop to zero if the pair have not interacted for a while (though nothing prevents the connection from reappearing in the future). If i and j meets ($F_{ij}^t = 1$ or $G_{ij}^t = 1$), then W_{ij}^t has a Poisson distribution with mean equal to a multiple (γ_h) of their old connection strength. γ_h signifies the rate of weight increase as a result of the "effectiveness" of a meeting; if $\gamma_h > 1$, then the weight will in general increase. (The weight may also decrease under the Poisson distribution, a consequence perhaps of unhappy meetings.) If i and j do not meet, their mean weight will decrease with rate $\gamma_\ell < 1$. Thus

$$W_{ij}^t \mid W_{ij}^{t-1}, F_{ij}^t, G_{ij}^t, \gamma_h, \gamma_\ell \sim \begin{cases} \mathrm{Poi}(\gamma_h(W_{ij}^{t-1} + \epsilon)) & \text{if } F_{ij}^t = 1 \text{ or } G_{ij}^t = 1 \\ \mathrm{Poi}(\gamma_\ell W_{ij}^{t-1}) & \text{o.w.} \end{cases} \tag{5}$$

where $W_{ij}^{t-1} = 0$ by default for $(i,j) \notin \mathrm{TRIAD}^t$, and ϵ is a small positive constant that lifts the Poisson mean away from zero. As W_{ij}^{t-1} becomes large, γ_h and γ_ℓ control the increase and decrease rates, and the effect of ϵ diminishes. γ_h and γ_ℓ have conjugate gamma priors:

$$\gamma_h \sim \mathrm{Gamma}(c_h, d_h), \tag{6}$$

$$\gamma_\ell \sim \mathrm{Gamma}(c_\ell, d_\ell). \tag{7}$$

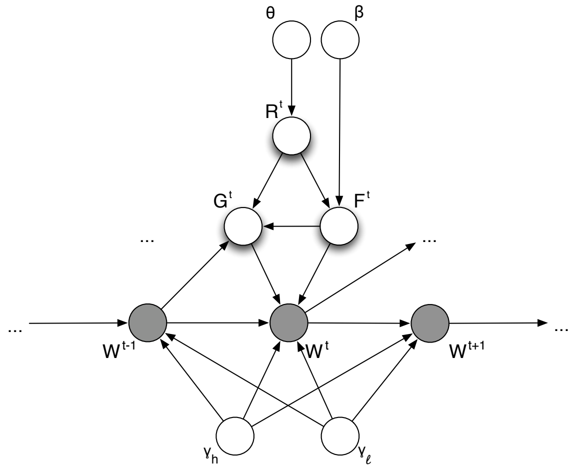

Fig. 1. Graphical representation of one time step of the generative model. R^t is a N_t-dimensional vector indicating each person's context at time t. F^t is a $N_t \times N_t$ matrix indicating pairwise dyadic meetings. G^t is a $N_{t-1} \times M_t$ matrix that indicate triadic closure for newcomers at time t. W^t is the matrix of observed connection weights at time t. θ, β, γ_h, and γ_ℓ are parameters of the model (hyperparameters are not shown).

Figure 1 contains a graphical representation of our model. The complete joint probability is:

$$\begin{aligned} P(\boldsymbol{\theta}, \boldsymbol{\beta}, \gamma_h, \gamma_\ell, W^{1:T}, R^{1:T}, F^{1:T}, G^{1:T}) = \\ P(\boldsymbol{\theta})P(\boldsymbol{\beta})P(\gamma_h)P(\gamma_\ell) \prod_t P(R^t|\boldsymbol{\theta})P(F^t|R^t, \boldsymbol{\beta}) \times \\ P(G^t|R^t, F^t, W^{t-1})P(W^t|G^t, F^t, W^{t-1}) \end{aligned} \quad (8)$$

3 Experiments

We illustrate the behavior of our model under different parameter settings on a set of established metrics.

3.1 Metrics

Degree distribution:
In an undirected graph, the degree of a node is its number of neighbors. For node i, we define its degree d_i to be $\sum_{j=1}^{N} I_{(W_{ij}>0)}$, and the average degree of the graph $\sum_{i=1}^{N} d_i/N$.

Node degrees in large natural networks often follow a power law distribution [5], i.e., the number of nodes D with degree n roughly conforms to the function $D(n) = n^{-\rho}$ for some exponent ρ. The value of ρ may vary from network to network, but the overall functional form remains the same. Intuitively, this means that there are many people with a few friends, and very few people with a lot of friends.

Clustering coefficient:
Across different social networks, it has often been observed that subsets of people tend to form fully-connected cliques. This inherent clustering tendency may be quantified by the *clustering coefficient* [6]. For the i-th node, C_i is defined to be the ratio between the number of edges E_i that actually exist between its d_i neighbors and the number of edges that would exist if the neighbors form a clique: $C_i = \frac{2E_i}{d_i(d_i-1)}$. The clustering coefficient of the whole network is the average over all nodes: $C = \sum_i C_i/N$.

Average path length:
We compute the length of the shortest path s_{ij} between every pair of nodes i and j. If i and j are not connected, then $s_{ij} = \infty$. Let $S := \{(i,j) \mid s_{ij} < \infty\}$ be the set of connected pairs. The average path length of the graph is defined to be $\bar{s} := \sum_{(i,j)\in S} s_{ij}/|S|$.

Effective diameter:
The diameter of a graph is the maximum of the shortest path distances between any pair of nodes: $\max_{(i,j)} s_{ij}$. If the graph consists of several disconnected clusters, its diameter is defined to be the maximum over all cluster diameters. Graph diameter can be heavily influenced by outliers. A more robust quantity is the effective diameter, commonly defined as the ninetieth percentile of all shortest paths. Let $\sigma(x)$ be the empirical quantile function of shortest path lengths, i.e., $\sigma(x) = \operatorname{argmax}_s\{s \mid f(s) < x\}$, where $f(s) = |\{(i,j) : s_{ij} < s\}|/N^2$ is the empirical cumulative distribution of s_{ij}. The effective diameter is taken to be $\sigma(.90)$, linearly interpolated if there is no exact match for the ninetieth percentile.

3.2 Simulations

We analyze the behavior of the model under different parameter settings using the four metrics introduced above. Albert and Barabási [5] and Newman [4] observe a wide range of values for these metrics in a variety of real social networks. Our model can generate networks whose clustering coefficient, average path length, and effective diameter fall within the range of observed values. Here we discuss how different parameter settings affect the values of these metrics, and provide intuition about why this is so.

Unless otherwise specified, the number of contexts K is set to 10. The context preference parameter θ_i is drawn from a peaked Dirichlet prior, where $\alpha_{k^*} = 5$ for a randomly selected k^*, and $\alpha_k = 1$ otherwise. This means that each person in the network has a slight preference for one context. The friendliness parameter β_i is drawn from a Beta(a, b) distribution, where $a = 1$ and b varies. The weights update rates are $\gamma_h = 2$, $\gamma_\ell = 0.5$, and $\epsilon = 1$. We add one person to the network at every time step, so that $n_t = t$. All experiments are repeated with 10 trials.

Friendliness The parameter β_i determines the "friendliness" of the i-th person and is drawn from a Beta(a, b) distribution. As b increases from 2 to 10, average friendliness decreases from 0.33 to 0.09. We wish to test the effect of b on overall network properties. In order to isolate the effects of friendliness, we fix the context assignments by setting $R_i^t = R_i^1$ for all $t > 1$. In this setting, people do not form triadic closures, and connection weights are updated only through dyadic meetings.

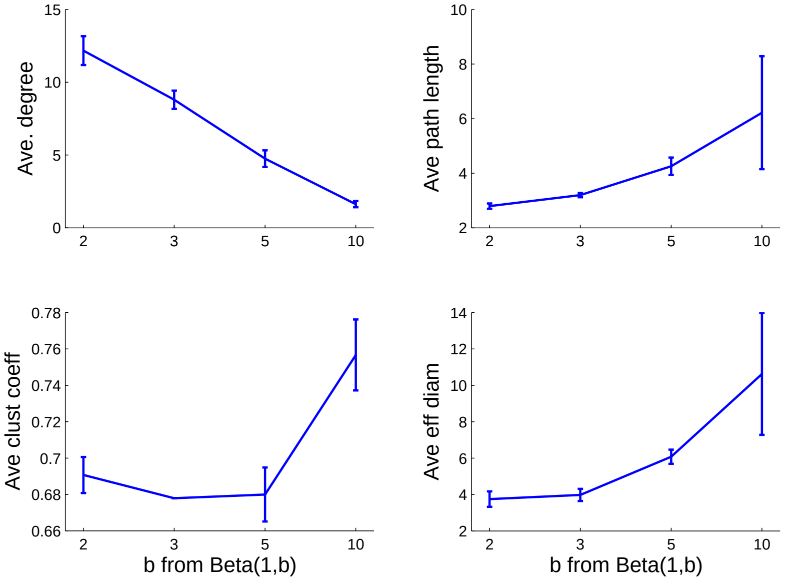

Fig. 2. Effects of the friendliness parameter on a network of 200 people with fixed contexts. The x-axes represent different values of b in $\text{Beta}(1, b)$.

As people become less friendly, one expects a corresponding decrease in average node degree. This is indeed what we observe in the average degree plot in Figure 2. Interestingly, the clustering coefficient goes up as friendliness goes down. This is because low friendliness makes for smaller clusters, and it is easier for smaller clusters to become densely connected than it is for bigger clusters. We also observe large variance in average path length and effective diameter at low friendliness levels. This is due to the fact that most clusters now contain one to two people. As small clusters become connected by chance, shortest path lengths varies from trial to trial.

Frequency of context switching In the current model, each person draws a new context at every time step. However, we can easily imagine a person working on one project for a while and then switching to the next project. When context switching is infrequent, people may develop stronger (and more) within-context relations.

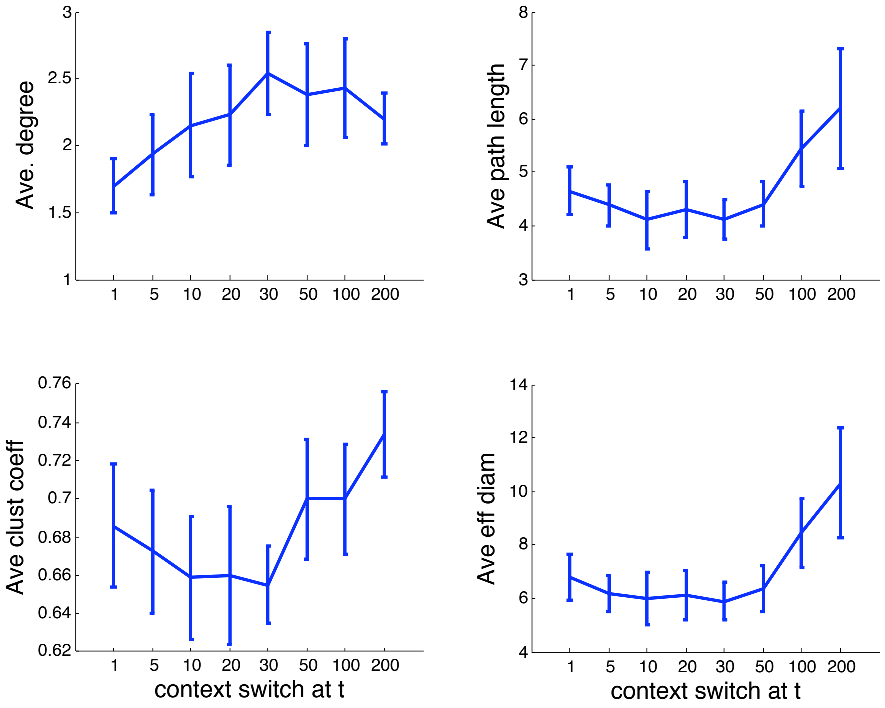

Fig. 3. Effects of the frequency of context switching on a network of 200 people. ($b = 8$)

We vary the frequency of context switching from 1 to 200 on a 200 node network. When the frequency is 1, people switch context at every time step; when the frequency is 200, contexts are fixed once and for all. In Figure 3, there appears to be a phase transition when context switching occurs every 30 time steps. This occurs as the consequence of two effects. First, when people switch contexts too frequently, they do not have the opportunity to meet everybody in the same context before moving on. Thus they have fewer neighbors and form smaller clusters on average. (As previously discussed, smaller clusters can lead to higher clustering coefficients.) Consequently, the average path length and effective diameter are also slightly long. On the other hand, when people never switch contexts (right-hand end of the x-axes), the number of neighbors is upper bounded by the number of people in the context. Clustering coefficient is high because everybody in the same context knows everybody else, and average path length and diamter are long because there are few paths to people outside of the current context.

Degree distribution Under different parameter settings, our model may generate networks with a variety of degree distributions. Lower levels of friendliness typically lead to more power-law-like degree distributions, while higher levels often result in a heavier tail. In Figure 4, we show two degree distribution plots for different friendliness levels. In the left-hand side plot, the quadratic polynomial is a much better fit than the linear one. This means that, when people are more friendly, the drop off in the number of people with high node degree is slower than would be expected under the power law. We do observe the power law effect at a lower level of friendliness. In the right-hand side plot, the linear polynomial with coefficient -1.6 gives as good of a fit as a quadratic function. This coefficient value lies well within the normally observed range for real social networks [5].

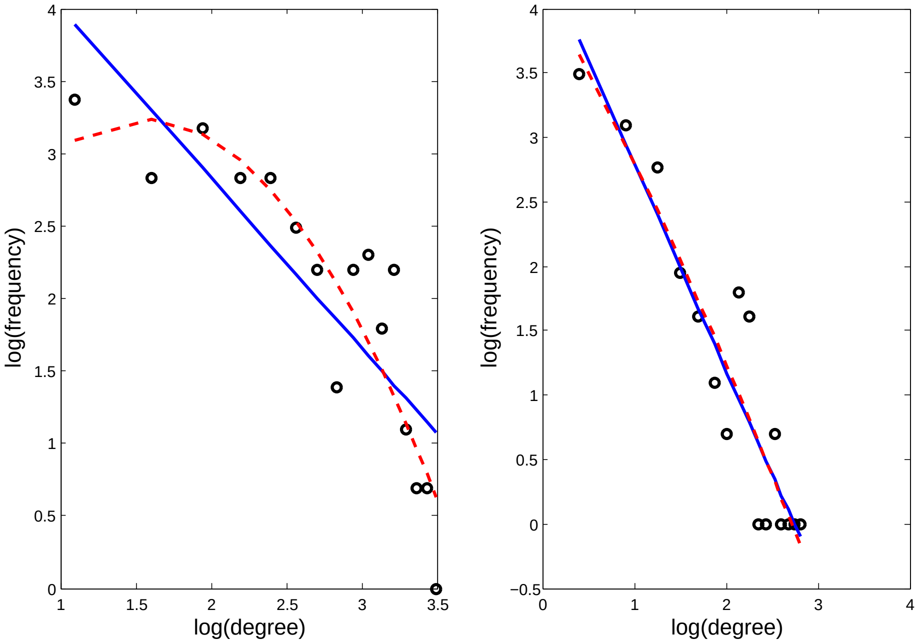

Fig. 4. Log-log plot of the degree distributions of a network with 200 people. β_i is drawn from Beta$(1, 3)$ for the plot on the left, and from Beta$(1, 8)$ for the right hand side. Solid lines represent a linear fit and dashed lines quadratic fit to the data. Contexts are drawn every 50 iterations.

Birth and death of links Our proposed model attempts to capture the dynamics of the birth and death of links. A link is born when the connection weight becomes non-zero, and the link dies when the weight returns to zero. Figure 5 shows link birth rates as the proportion of newly established ties to the number of possible births, and link death rates as the proportion of the number of deaths to the number of links that exist at that point in time.

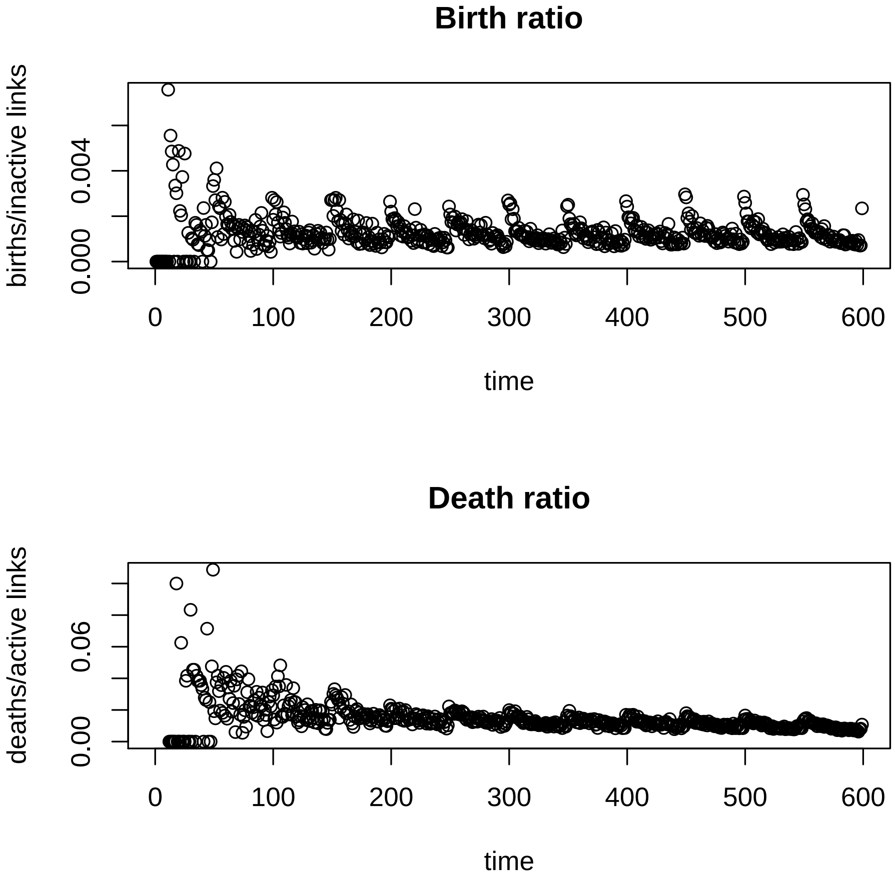

Fig. 5. Birth (top) and death (bottom) of links in a network of 600 people over 600 time steps. Contexts switches occur every 50 iterations, $K = 20$ and $b = 10$.

At the beginning, there are few existing links. Therefore the birth rate is relatively high. Since one person is added to the network at each time step, the number of possible connections grows as $t(t-1)/2$. Thus the birth rate becomes smaller at larger values of t. We note periodical trends in both births and deaths of links. This periodicity coincides with changes in context. At each context switch, a fresh pool of possible connections becomes available, and weaker links from previous connections are now more likely to die out.

Weight distributions One of the main strengths of our model lies in its ability to represent weighted links. In real life, friendships are not simply existent or absent. A strong connection should take longer to dissipate than would a weak connection. Link weights act as memory in preserving friendships. Old friendships may be rekindled if the pair rotate within similar contexts. We compare the evolution of simulated weights with email exchange in the well-known Enron dataset. Figure 6 shows typical weight progressions over time in a simulated network. Figure 7 plots typical patterns of weekly email exchange counts between Enron employees. Our model is clearly capable of reproducing both long-lasting and short-range connections. Previously severed links can be renewed, as is the case for the pair $(45, 47)$.

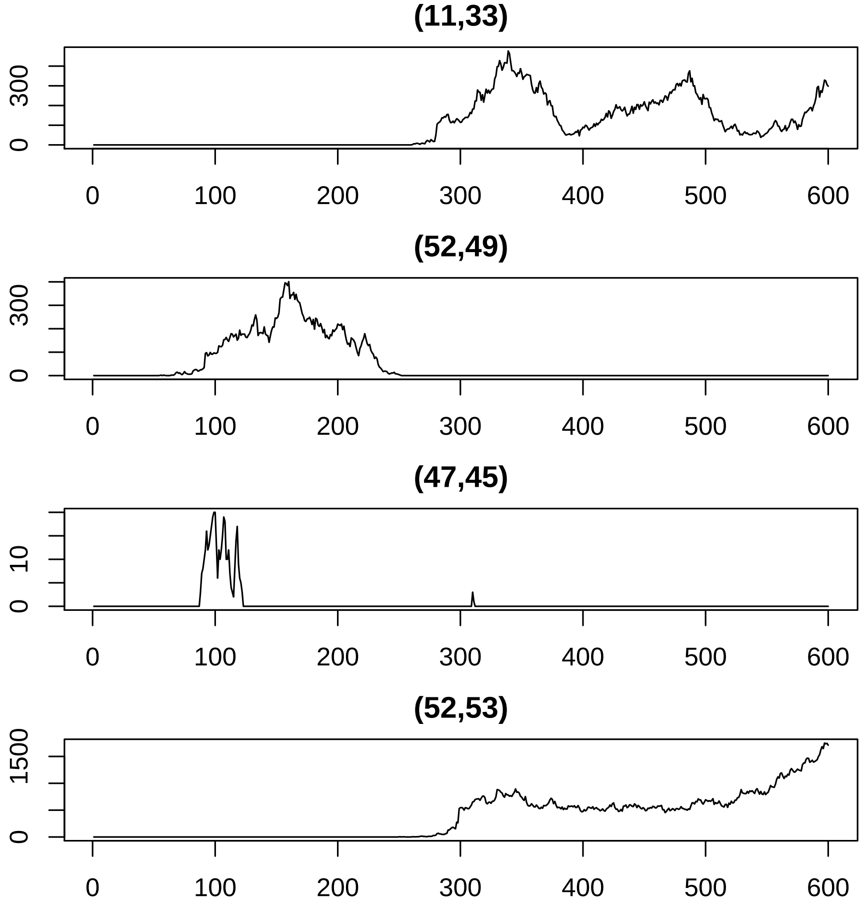

Fig. 6. Weight dynamics for 4 different pairs in a network of 600 people over 600 time steps. Contexts switches occur every 50 iterations and $b = 3$.

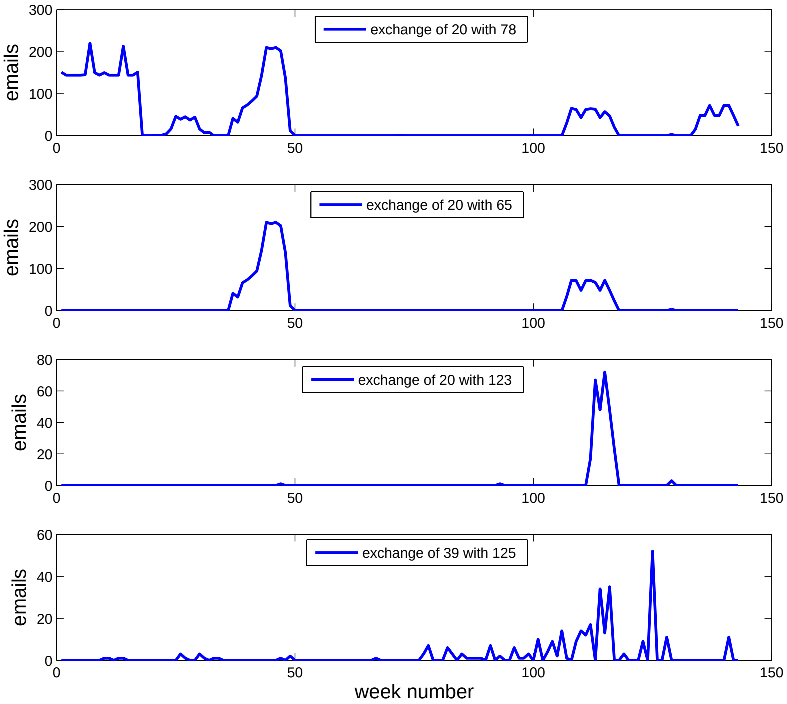

Fig. 7. Weekly email exchange counts for four randomly selected pairs between 136 Enron employees.

4 Learning Parameters via Gibbs Sampling

Parameter learning in DCFM is possible via Gibbs sampling. We leave a detailed investigation of learning results to another paper, but give the Gibbs updates here for reference. Using ... as a shorthand for "all other variables in the model," we have:

$$\boldsymbol{\theta}_i \mid \ldots \sim \mathrm{Dir}(\boldsymbol{\alpha} + \boldsymbol{\alpha}'_i), \tag{9}$$

$$P(\beta_i \mid \ldots) \propto \beta_i^{A_i+a-1}(1-\beta_i)^{b-1} \prod_{j \neq i} (1 - \beta_i \beta_j)^{B_{ij}}, \tag{10}$$

$$\gamma_h \mid \ldots \sim \mathrm{Gamma}(c_h + w_h, (v_h + 1/d_h)^{-1}), \tag{11}$$

$$\gamma_\ell \mid \ldots \sim \mathrm{Gamma}(c_\ell + w_\ell, (v_\ell + 1/d_\ell)^{-1}). \tag{12}$$

In Equation 9, $\alpha'_{ik} := \sum_{t=1}^{T} I_{(R_i=k)}$ is the total number of times person i is seen in context k. In Equation 10, $A_i := |\{(j,t) \mid R_i^t = R_j^t \text{ and } F_{ij}^t = 1\}|$ is the total number of dyadic meetings between i and any other person, and $B_{ij} := |\{t \mid R_i^t = R_j^t \text{ and } F_{ij}^t = 0\}|$ is the total number of times i has "missed" an opportunity for a dyadic meeting. Let $H := \{(i,j,t) \mid F_{ij}^t = 1 \text{ or } G_{ij} = 1\}$ represent the union of the set of dyadic and triadic meetings, and $\mathcal{L} := \{(i,j,t) \mid (i,j) \in \mathrm{DYAD}^t \text{ and } F_{ij}^t = 0\}$ the set of missed dyadic meeting opportunities. $w_h := \sum_{(i,j,t) \in \mathcal{H}} W_{ij}^t$ is the sum of updated weights after the meetings, and $v_h :=$

$\sum_{(i,j,t)\in\mathcal{H}}(W_{ij}^{t-1}+\epsilon)$ is the sum of the original weights plus a fixed constant. $w_l := \sum_{(i,j,t)\in\mathcal{L}} W_{ij}^t$ is the sum of weights after the missed meetings, and $v_l := \sum_{(i,j,t)\in\mathcal{L}} W_{ij}^{t-1}$ is the sum of original weights. (Here we use zero as the default value for W_{ij}^{t-1} if j is not yet present in the network at time $t-1$.)

Due to coupling from the pairwise interaction terms $\beta_i\beta_j$, the posterior probability distribution of β_i cannot be written in a closed form. However, since β_i lies in the range $[0,1]$, one can perform coarse-scale numerical integration and sample from interpolated histograms. Alternatively, one can design Metropolis-Hasting updates for β_i, which has the advantage of maintaining a proper Markov chain.

The variables F_{ij}^t and G_{ij} are conditionally dependent given the observed weight matrices. If a pairwise connection W_{ij} increases from zero to a positive value at time t, then i and j must either have a dyadic or a triadic meeting. On the other hand, dyadic meetings are possible only when i and j are in the same context, and triadic meetings when they are in different contexts. Hence F_{ij}^t and G_{ij}^t may never both be 1. In order to ensure consistency, F_{ij}^t and G_{ij} must be updated together. For $(i,j) \in \mathrm{TRIAD}^t$,

$$\begin{aligned}
&P(F_{ij}^t = 1, G_{ij} = 0 \mid \ldots) \propto I_{(R_i^t = R_j^t)}(\beta_i\beta_j)\mathrm{Poi}(W_{ij}^t; \gamma_h\epsilon),\\
&P(F_{ij}^t = 0, G_{ij} = 1 \mid \ldots) \propto I_{(R_i^t \neq R_j^t)}\mu_{ij}\mathrm{Poi}(W_{ij}^t; \gamma_h\epsilon),\\
&P(F_{ij}^t = 0, G_{ij} = 0 \mid \ldots) \propto \left[I_{(R_i^t = R_j^t)}(1-\beta_i\beta_j) + I_{(R_i^t \neq R_j^t)}(1-\mu_{ij})\right] I_{(W_{ij}^t = 0)}.
\end{aligned} \tag{13}$$

For $(i,j) \in \mathrm{DYAD}^t \backslash \mathrm{TRIAD}^t$,

$$\begin{aligned}
&P(F_{ij}^t = 1 \mid \ldots) \propto I_{(R_i^t = R_j^t)}(\beta_i\beta_j)\mathrm{Poi}(W_{ij}^t; \gamma_h(W_{ij}^{t-1}+\epsilon)),\\
&P(F_{ij}^t = 0 \mid \ldots) \propto (I_{(R_i^t = R_j^t)}(1-\beta_i\beta_j) + I_{(R_i^t \neq R_j^t)})\mathrm{Poi}(W_{ij}^t; \gamma_\ell W_{ij}^{t-1}).
\end{aligned} \tag{14}$$

There are also consistency constraints for R^t. For example, if $F_{ij}^t = F_{jk}^t = 1$, then i, j, and k must all lie within the same context. If $G_{kl} = 1$ in addition, then l must belong to a different context from i, j, and k. The F variables propagate transitivity constraints, whereas G propagates exclusion constraints.

To update R^t, we first find connected components within F^t. Let p denote the number of components and I the index set for the nodes in the i-th component. We update each R_I^t as a block. Imagine an auxiliary graph where nodes represent these connected components and edges represent exclusion constraints specified by G, i.e., I is connected to J if $G_{ij} = 1$ for some $i \in I$ and $j \in J$. Finding a consistent setting for R^t is equivalent to finding a feasible K-coloring of the auxiliary graph, where K is the total number of contexts. We sample R_I^t sequentially according to an arbitrary ordering of the components. Let $\pi(I)$ denote the set of components that are updated before I. The posterior probabilities are:

$$P(R_I^t = k \mid R_{\pi(I)}^t, G) \propto \begin{cases} 0 & \text{if } G_{IJ} = 1 \text{ and } R_J^t = k \text{ for some } J \in \pi(I) \\ \prod_{i \in I} \theta_{ik} & \text{o.w.} \end{cases} \tag{15}$$

These sequential updates correspond to a greedy K-coloring algorithm; they are approximate Gibbs sampling steps in the sense that they do not condition on the entire set of connected components.

4.1 Parameter Learning Results

As a sanity check, we apply the parameter learning algorithm to a network generated from the model. The hyperparameters are set to be those used in the simulation, though in practice we find the Gibbs sampler to be robust to changes in hyperparameter settings. The learning procedure quickly converges to a stable set of parameter values. Below are convergence plots for γ_h, γ_ℓ and β for a small dataset of 78 people in 10 contexts where links form and dissolve over 84 time steps. The number of Gibbs iterations is $10,000$.

Figure 8 contains a scatter plot of the friendliness β parameters (mean of the posterior vs. true values). Figure 9 contains the convergence plot and the posterior distribution of γ_h and γ_ℓ. Note that, due to noise in the sampling process, the γ_h values oscillate around the median of 1.90 (true value being 2) and γ_ℓ values have median 0.96 (true value being 1). We observe similar convergence trends for θ_i and omit the figure due to lack of space.

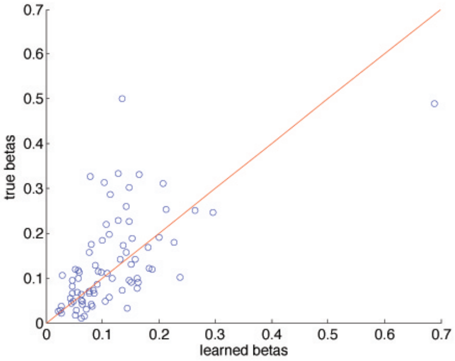

Fig. 8. β parameter scatter plot.

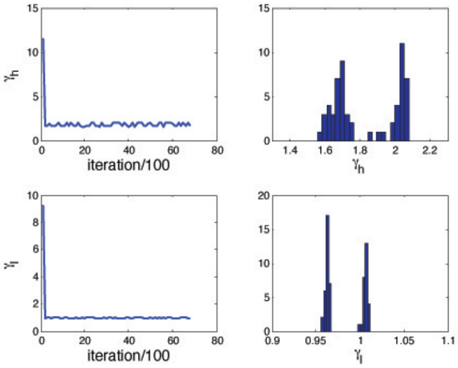

Fig. 9. The left-hand column shows convergence plots of γ_h and γ_ℓ over Gibbs sampling iterations. The right-hand column contains histogram of the sampled values.

To test the interpretability of the model on real data, we learn the parameters of DCFM on a real-life collaboration network. We have collected interaction information, such as meetings and co-authorships, over 13 years for 120 people connected to our lab. We omit the names of people due to anonymity constraints. More information will be available in the full version of the paper.

Sampling results yield the median values of $\gamma_h = 1$ and $\gamma_l = .02$. This shows that lab members either have steady collaboration patterns over time, or have spurious interactions that quickly die off. We also find that the head of the lab, who participates in most but not all of the collaborations, is quite friendly with $\beta = .86$. Interestingly, the next most friendly person with $\beta = .82$ is a student who is not the most prolific but has many co-authors. In the process of parameter learning, we find that our original assumption of 10 contexts is not enough to accommodate all the consistency constraints arising between R, F, and G. Thus we increase the number of contexts to 20. Figure 10 shows the learned context distributions of the above mentioned professor and student. The two are mostly comparable except for contexts 7, 8, and 17. Assuming that the contexts represent topics of study, the student is the most interested in 7 and least in 17, whereas the professor has a rather uniform distribution over all fields, most of all number 8.

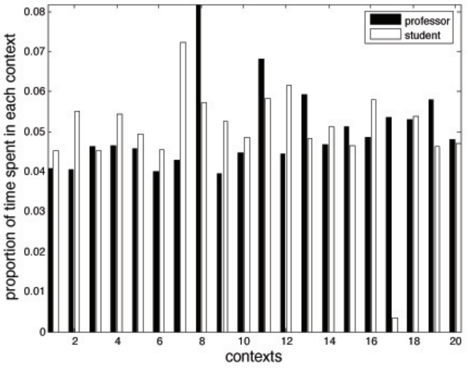

Fig. 10. Context distributions of the two most "friendly" people in the co-authorship network.

5 Possible Extensions

5.1 Evolution of Context Preferences

A person's context distribution is influenced by the social groups to which he belongs. People who are friends with gym-goers may start to frequent the gym themselves. Thus it could be desirable to incorporate evolution of the θ parameters (indicating context preference) into our model. We propose to update θ for each person using the θ parameters of his neighbors, weighted by the connection strengths:

$$\theta_i^t = \lambda \theta_i^{t-1} + (1-\lambda) \frac{1}{\sum_j W_{ij}^t} \sum_j W_{ij}^t \theta_j^{t-1}. \tag{16}$$

The larger λ (a person's independence) is, the less susceptible the person is to the preference of his friends.

5.2 Long Term Memory

Weighted links capture the effect of short term memory; in our model, a link established at time t will likely remain at time $t+1$. However, once the weight becomes zero, renewal of the link becomes is likely as a 'birth' of a new link. To capture long term memory, we could model weights as a continuous gamma distribution, so that established links always carry small residual weights. The drawback is that the weight matrices will be dense, and we would need an additional thresholding parameter for the 'death' of a link. Alternatively, at the cost of introducing N new parameters, we can make each person 'remember' the strength and duration of his past connections.

6 Related Work

The principles underlying the mechanisms by which relationships evolve are still not well understood [7]. Current models aim at either describing observed phenomena or predicting future trends. A common approach is to select a set of graph based features, such as degree distribution or the number of dyads and triangles, and create models that mimic observed behavior of the evolution of these features in real life networks. Works in physics [8, 9, 10] and in social sciences [11, 12] follow this approach. However, under models of average behavior, the actual links between any two given people might not have any meaning. Consequently, these models are often difficult to interpret.

Another approach aims to predict future friends and collaborators based on the properties of the network seen so far [4, 7]. These models often cannot encode common network dynamics such as mobility and link modification. Moreover, these models usually do not take into account triadic closure, a phenomenon of great importance in social networks [2, 13].

In [14], Sarkar and Moore present an interesting dynamic social network model (with fixed number of people). This work builds upon a previous model by Hoff, Raftery, and Handcock [15], which introduces latent positions for each person in order to explain observed links. If two people are close in the latent space, they are likely to have a connection. Hoff, Raftery, and Handcock estimate latent positions in a static data set, whereas Sarkar and Moore add a dynamic component by allowing the latent positions to be updated based on both their previous positions and on the newly observed interactions. One can imagine a generative mechanism that governs such perturbations of latent positions. In fact, the DCFM model presented in this paper can be seen as a generative model for the latent mapping function.

7 Discussion

Our focus on generative modeling in this paper is prompted by the need to provide a plausible explanation for how networks form and evolve. It is flexible and can be adapted to alternative theories of the friend evolution process. For example, in our model, the decision to allow links to decay is made independently on each pair. However, theory of Simmelian ties [16] suggest that two people who are no longer friends may nevertheless remain so due to influence from a third party. This is a plausible alternative to our current model.

Our choice of modeling weighted networks is motivated by the fact that friednships between people are not binary. Stronger links tend to last longer periods of time; temporary connections cease to exist once the cause disappears. However, it is often difficult to obtain real datasets with weighted connections. We propose to use the number of email, sms and phone call exchanges in preset time intervals as a proxy to the weight of links between people. This is a very coarse representation of a relationship weight, since non-communication does not necessarily imply change in link weight. Hence the DCFM model may predict smoother connection weights than the observed values.

To show that our model is capable of generating realistic social environments, we provide simulation results that adhere to observations made on realistic datasets in [17]. However, there is no groundtruth for the parameters in the hidden layer. Variables that address context choice and meeting occurrance at time step t have to be inferred from the previous and currently observed weights alone. This brings up the question of identifiability. Unfortunately, the complexity of the model makes it difficult to answer this question and we are currently exploring possible solutions to this problem.

Another interesting question is exchangeability. The earlier a person appears in the network, the more chances he has to establish connections. People who have been in the network longer are expected to have more connections and thus nodes (people) are not exchangeable over time.

The current model does not place any explicit upper bounds on the number of links a person can establish. It is effectively limited by the number of people in the same context. Unless a person is very friendly and has uniform distribution, the number of links is not expected to be high. In realistic networks, we expect the context preference distribution and friendliness to be skewed, because a person has a limited amount of time and energy to build and maintain relationships.

In conclusion, we provide an exploratory study of a new generative model for dynamic social networks in this paper. Simulation results demonstrate the advantages as well as shortcomings of this model. In future work, we hope to address issues of identifiability and investigate possible extensions of this work.

References

[1] Milgram, S.: The small-world problem. Psychology Today (1967)

[2] Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Cambridge University Press, Cambridge (1994)

[3] Barabási, A.L., Albert, R.: Emergence of scaling in random networks. Science **286** (1999) 509–512

[4] Newman, M.: The structure of scientific collaboration networks. In: Proceedings of the National Academy of Sciences USA. Volume 98. (2001) 404–409

[5] Albert, R., Barabási, A.: Statistical mechanics of social networks. Rev of Modern Physics **74** (2002)

[6] Watts, D., Strogatz, S.: Collective dynamics of ”smallworld” networks. Nature **393** (1998) 440–442

[7] Liben-Nowell, D., Kleinberg, J.: The link prediction problem for social networks. In: Proc. 12th International Conference on Information and Knowledge Management. (2003)

[8] Jin, E., Girvan, M., Newman, M.: The structure of growing social networks. Physical Review Letters E **64** (2001)

[9] Barabasi, A., Jeong, H., Neda, Z., Ravasz, E., Schubert, A., Vicsek, T.: Evolution of the social network of scientific collaboration. Physica A **311**(3–4) (2002) 590–614

[10] Davidsen, J., Ebel, J., Bornholdt, S.: Emergence of a small world from local interactions: Modeling acquaintance networks. Physical Review Letters **88** (2002)

[11] Van De Bunt, G., Duijin, M.V., Snijders, T.: Friendship networks through time: An actor-oriented dynamic statistical network model. Computation and Mathematical Organization Theory **5**(2) (1999) 167–192

[12] Huisman, M., Snijders, T.: Statistical analysis of longitudinal network data with changing composition. Sociological Methods and Research **32**(2) (2003) 253–287

[13] Kossinets, G., Watts, D.: Empirical analysis of an evolving social network. Science **311**(5757) (2006) 88–90

[14] Sarkar, P., Moore, A.: Dynamic social network analysis using latent space models. SIGKDD Explorations: Special Edition on Link Mining (2005)

[15] Hoff, P., Raftery, A., Handcock, M.: Latent space approaches to social network analysis. Journal of the American Statistical Association **97** (2002) 1090–1098

[16] Krackhardt, D.: The ties that torture: Simmelian tie analysis in organizations. Research in the Sociology of Organizations (1999)

[17] Albert, R., Barabási, A.L.: Dynamics of complex systems: Scaling laws for the period of boolean networks. Physical Review Letters **84** (2000) 5660–5663