Extending Generative Models of Large Scale Networks

Eli Stickgold
Bruce Skarin
Ian Stewart
Corey Lofdahl
Charles River Analytics
625 Mt. Auburn Street
Cambridge, MA 02138
617-491-3474
estickgold@cra.com, bruceskarin@gmail.com, istewart@cra.com, clofdahl@cra.com

ABSTRACT: Since the launch of Facebook in 2004 and Twitter in 2006, the amount of publicly available social network data has grown in both scale and complexity. This growth presents significant challenges to conventional network analysis methods that rely primarily on structure. In this paper, we describe a generative model that extends structure-based connection preference methods to include preferences based on agent similarity or homophily. We also discuss novel methods for extracting model parameters from existing large scale networks (e.g. Twitter) to improve model accuracy. We demonstrate the validity of our proposed extensions and parameter extraction methods by comparing model-generated networks with and without the extensions to real-life networks based on metrics for both structure and homophily. Finally we discuss the potential implications for including homophily in models of social networks and information propagation.

ACKNOWLEDGEMENTS: This work was performed under DARPA contract number W31P4Q-12-C-0235. The authors thank Dr. Rand Waltzman for his significant technical support and eager engagement on this project. This work was funded in its entirety by the Information Innovation Office (I20). The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government. We also acknowledge the contribution of Professor Frank Witmer who helped with the initial modeling efforts.

1. Introduction

The proliferation of social media has profoundly changed the information landscape since the early 2000s. The emergence of massive social media services such as Facebook and Twitter (Java, Song, Finin et al., 2007; Viswanath, Mislove, Cha et al., 2009) has led to the proliferation of massive, highly-connected social networks. Understanding modern age information flows requires understanding the way these networks grow, evolve and decay over time (Hughes, Rowe, Batey et al., 2012). By their nature however, large portions of these networks are invisible, either hidden by privacy policies or existing in entirely different media: a strong link may exist between two real-world friends who rarely interact directly on Twitter, but nevertheless have a strong impact on each other. This creates ethical risks and technical challenges for direct study, even though most hidden users are often qualitatively similar to visible users (Madden, 2012).

Early analysis attempts solved the first problem by using anonymized datasets, which left the structural information intact, but removed the personally identifying information. Unfortunately, it was quickly proved that even sparse anonymized networks are vulnerable to de-anonymizing attacks with very little information required (Narayanan & Shmatikov, 2008). Furthermore, completely anonymized networks have the problem of removing vital context regarding the actors and links in the network.

Given these risks and challenges, simulation models, and in particular agent based models, can help provide a better analogue for study and research. More realistic models also have the potential to improve our ability to understand and predict information propagation in modern social networks that can only be partially observed. Furthermore, while the exact meaning of links within a social network may be debated (e.g. a personal
friendship is different from a fan-celebrity relationship),

virtually all social networks include information-diffusion
aspects that can in turn be modeled.

In this paper, we describe an agent based model that
leverages social science theory and existing network
modeling methods to produce synthetic directed,
weighted social networks. This model is then used in a

case study focused on Twitter networks, which provides
both model parameters and a ground-truth for conducting
network comparisons. Such a model may offer insight to
analysts seeking patterns in link-formation between users,
particularly in cases of news-related information
diffusion.

We begin by explaining the need for features based on
social science and how they affect the process of agents
forming and updating. In particular, we focus on the
principle of homophily (McPherson, Smith-Lovin, Cook
et al., 2001), which has been found to have strong
implications on the spread and interpretation of
information within a network. We then describe our
model for reproducing homophily in synthetic networks.
We follow the model with a discussion of the novel
methods we used for extracting parameters from existing

network data. In closing we compare our models with and
without the new features to existing Twitter networks and
discuss the implications for further research and
application. In summary, our findings indicate that in the
context of representing social networks, the addition of
these new features provides measurable improvements
over other network generation methods.

2. Structure and Homophily

There already are a variety of methods for generating
networks of different structural archetypes like small-
world, random, and scale-free. Some common approaches
include the Watts-Strogatz, Erdős–Rényi (ER), and
Barabási–Albert (Barabási & Albert, 1999) models.
These have proven to be accurate analogs for the structure
of many real-world networks using metrics such as the
degree distribution, diameter, and clustering coefficient.

Yet for some networks, in addition to structural features
there are also social and functional features that affect

how networks create additional archetypes such as the
polarized crowd, community clustered, customer support,
and broadcast networks as described in Smith, Rainie,
Shneiderman & Himelboim (2014). In cases such as these,
a different approach is required to represent the
basis for forming social/functional features in network
models. In addition, standard statistical models such as
the ER random graph generator are often inadequate in
replicating the structure of social networks (Mislove,
Marcon, Gummadi, Druschel, & Bhattacharjee, 2007),
which inspires the search for more accurate models.

Research into more robust methods of modeling social
networks includes latent space models, Hidden Markov
Models, and actor-oriented models (Snijders, 2011). One
advantage of actor-oriented models is the ability to
include behavior-based theory instead of relying on a
purely statistical approach. Actor-oriented models also
enable more natural extensions to the modeling of other
network behaviors such as information propagation and
network adaptation and evolution.

In focusing on behaviors related to social networks, the
principle of homophily provides a useful heuristic for
determining the likelihood of link formation between two
given actors. Homophily asserts that similar people
interact more frequently than dissimilar people. In cases
such as language, the assertion of homophily holds
strongly (Hale, 2014), while in other cases there can be a
wider degree of variation. Homophily has also been found
to have powerful implications on the information people
receive and the attitudes they form. For these reasons,
homophily is an important feature for generating realistic
analogues of existing social networks.

3. Homophily in Agent-Based Modeling

Agent-based modeling is an ideal formalism for a
behavior-based approach to the generation of social
networks because of its flexibility in representing diverse
attributes and behaviors. In an agent based model,
individual actors can incorporate a dynamic range of
demographic and social attributes and behaviors for use in
determining how links are formed between actors.

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Example Entities/Range</th>
<th>Use</th>
<th>Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language</td>
<td>Categorical</td>
<td>English, Spanish, French, etc.</td>
<td>Both</td>
<td>High</td>
</tr>
<tr>
<td>Gender</td>
<td>Categorical</td>
<td>Male, Female</td>
<td>Link formation</td>
<td>Moderate</td>
</tr>
<tr>
<td>Activity</td>
<td>Ordinal</td>
<td>High, Medium, Low</td>
<td>Link formation</td>
<td>Low</td>
</tr>
<tr>
<td>Status</td>
<td>Scalar</td>
<td>Power law distribution</td>
<td>Messaging</td>
<td>Moderate</td>
</tr>
</tbody>
</table>
Demographic types that might be represented include categorical (e.g. gender, language), ordinal (e.g. time zone, ranking), and scalar (e.g. latitude/longitude). Each demographic can be parameterized with a distribution for the number of members, and by ranking their importance to link formation. Table 1 provides several example demographics that were used in our research. To determine if a link is formed between two given agents we used a modified version of the preference function described in Pasta, Jan, Zaidi, & Rozenblat (2013).

After a number of experiments generating networks of different sizes, we also found the best fit when we applied the function to different subgroups at each stage of network formation. For initialization the preference function is applied to all of possible initial agent pairings. During early growth, after the seed network is created, the preference function is then applied only to a selected subgroup (e.g. language, gender, etc.) for a given agent. After an agent has a network larger than a specified degree, then the function is applied only to connections within a given degree distance.

3.1 Modified Preference Function

To recreate the principle of homophily, each agent uses a modified version of the preference function commonly used in structural approaches to network generation. The algorithm used here builds on the work of Pasta, Jan, Zaidi, & Rozenblat (2013) by taking a weighted average of the similarity between two given agents based on the relative importance of each demographic. For each demographic $D_p$, of the type categorical, the similarity for any two agents $i$ and $j$ is assigned as such:

$$D_p(i, j) = \begin{cases} 1 & \text{if } i_p = j_p \\ 0 & \text{if } i_p \neq j_p \end{cases}$$

For ordinal and scalar demographic types, the similarity between two agents is simply the normalized Manhattan distance, where $\delta_p$ represents the difference between the maximum and minimum values of the range that can be assigned to a given demographic $D_p$:

$$D_p(i, j) = 1 - \frac{|i_p - j_p|}{\delta_p}$$

The cumulative similarity $S$ between any two actors $i$ and $j$ is then the weighted arithmetic mean, calculated over $n$ demographics. Where $w$ is the weight assigned to each demographic regarding its importance in link formation:

$$S_{i,j} = \frac{\sum_{p=1}^{n} w_p D_p}{\sum_{p=1}^{n} w_p}$$

For the structural preference functions, we selected triadic closures (aka friend-of-a-friend) and degree preference. The triadic closure preference $T$ between two actors $i$ and $j$ is formulated as

$$T_{i,j} = \frac{|i \cap j|}{\min(|i|, |j|)}$$

where $\min(|i|, |j|)$ represents the minimum number of edges in either actor’s network. Using the minimum ensures that agents with a smaller number of edges are not penalized against agents with large networks. Degree preference $G$ is then assigned by the following:

$$G_{i,j} = \frac{\deg_j}{\max(\deg_n)}$$

With this formulation, the degree preference of $i$ is based on the out-degree of $j$ relative to the maximum out-degree within the entire network. If $j$ is the agent with the most followers, it will receive a score of 1.

Similar to the cumulative similarity, the cumulative structural preference $R$ is then:

$$R_{i,j} = \frac{w_p T_{i,j} + w_p G_{i,j}}{\sum_{p=1}^{n} w_p}$$

Finally, the total preference $C$ for agent $i$ to follow actor $j$ is given by the weighted sum of the respective cumulative measures:

$$C_{i,j} = w_R R_{i,j} + w_S S_{i,j}$$

When an agent is selected to form a connection, this preference function is used to create a distribution of measurements that is then randomly sampled for a specified number of edges. As discussed in detail in Section 5, the resulting network is now able to match existing measures of structure (e.g. network diameter and degree distribution) as well as measures related to homophily (e.g. similarity and cohesion).

3.2 Two-stage Selection Process

Even though preliminary assessments found improvements in network similarity, we noticed that for certain demographic categories such as language, the modularity was more significant than for other demographic categories such as language, the modularity was much more significant than could be accounted for in the preference function. This becomes particularly apparent as the size of the network grows and the weight of any given features is diluted.
For this reason after initialization and during early network growth we applied a two-stage process. In the first stage, an agent determines from which demographic category to draw from. Then in the second stage the agent selects a link based on the preference function as applied only to the subgroup. This process is similar to other latent-variable models such as composite-network friendship detection (Zhong, Xiang, Fan, Liu & Yang 2014).

After an agent has a network larger than a given degree, then instead of using all the agents within a given demographic category the pool of possible connections is built from the connected components for a given degree distance. From a behavioral perspective this two stage filtering is more representative of the individual constraints for large social network mediums like Twitter. That is, when deciding who to follow, new users do no look at the whole of Twitter, but rather within certain categories like language and from the referrals within his or her existing network.

3.3 Network Generation

The majority of the parameters used for generating networks are extracted from an existing social network (as described in the following Section 4). The remainder are currently set using manual experimentation. Both agents and links are added to the network over time at a rate based on the average growth rate of the observed network. New agents are created as follows:

Input: Demographics, Parameters
Output: New agent
for each Demographic do
    P=random number between (0,1)
    D=Assign category/value Dn(P)
end for
Links are then formed by the following:
Input: AgentPool
Output: New link
Total=0
for agenti < AgentPool do
    for agentj < AgentPool do
        Total = Total + Ci,j
    end for
    P=random number between (0, Total)
    Source = AgentPool(C(P)
end for
else SourcePool=ConnectedComponent
Total=0
for agenti < SourcePool do
    Total = Total + C_target,i
end for
P=random number between (0, Total)
Source = SourcePool(C(P)

3.4 Extracting Model Parameters

In order to make use of these extensions, we needed a way to determine appropriate values for each parameter. Because these values depend on the network being modeled, we developed a technique for extracting values from a representative network that could be used to determine accurate values for any of a number of possible use-cases. To test our technique, several social networks were constructed from sets of tweets collected over the course of a week in September on a variety of news topics using Twitter’s API. The corpus for this study was created by filtering the public stream of tweets for keywords related to then-current news topics. From that corpus, we filtered tweets for the keyword “Ukraine” and collected approximately 650,000 tweets to use as our primary dataset. We chose this topic because of Ukraine’s prominence in international news at the time of the study, due to the nation’s public conflict with Russia. The relatively high volume and diversity of the participants involved ensured a reasonable ground-truth dataset.

Following collection, the team converted the raw tweets into a directed network using the @-symbol as a proxy for an edge between users. For example, the tweet “@userA breaking Ukraine news!” from user B would result in a directed edge from B to A. The team chose this tweet-based approach to social network construction over a typical snowball-sampling approach that selects one prominent network actor, collects all of their followers, and repeats the process on these new actors. We made this decision for two main reasons:

1. **Scalability**: it is easier to build a large network from tweets than query Twitter recursively for an ever-growing network.
2. **Information diffusion**: the use of the @-symbol implies diffusion of information from one user to another, which is a key area of interest for this study.

In addition to this basic structural information, we queried Twitter for user-specific data to provide node attributes to the network. In contrast to similar social media studies (De Choudhury, 2011), we only extracted self-reported attributes available directly through the API, rather than attempting to speculate on inaccessible attributes such as gender. This decision limits the overall amount of user
data available but more importantly ensures that the demographics for our model’s agents are based as much as possible in reality and less on prior expectations.

Once the networks were built, a set of static and dynamic metrics were extracted to quantify basic network trends. The key to choosing the metrics was generalizability across potential models, meaning statistics that can differentiate a range of known network types. For instance, the ratio of edges to nodes is higher in small-world networks than in scale-free (hierarchical) and random networks. The following metrics were extracted as initialization parameters for the model:

- **Actor pool size**: the total number of nodes in the network, intended as a cap to the model’s growth.
- **Nodes per hour**: number of new nodes that appear in the network in a given hour.
- **Edges per hour**: number of new edges formed in the network in an hour.
- **Edge adding rate**: the µ and σ of every node’s edge-adding rate, i.e. edges/node/hour.
- **Activity frequency**: the distribution of all users’ frequency of posting statuses, i.e. statuses/day.
- **Language split**: percentages of the languages spoken by users in the network. All languages with percentages below 5% were classified as “Other” to minimize data sparseness.

After extracting these parameters for model initialization, we extracted unweighted in-degree and out-degree as a means of comparing real and synthetic social networks. The in-degree of a node \( n \) is equal to the number of edges incident on \( n \), while out-degree is equal to the number of edges leaving \( n \). Using the overall distributions for in- and out-degree, we were able to establish quantitative goals for the model to achieve over the course of its simulation.

4. Results

Preliminary results demonstrate the success of incorporating homophily into the network generation process. We find that the model accounting for homophily in user demographics produces the most realistic network in terms of in- and out-degree (5.1). In addition, our model yields emergent properties to match the ground-truth network, including homophilic clustering (5.2).

For our preliminary results, we used a set of four networks of nearly identical scale: (1) a slice of the ground-truth (GT) network from tweets; (2) a synthetic scale-free network generated using the B model of the Barabási-Albert (BA) algorithm (Barabási and Albert 1999); (3) a network generated using only structural preference (S); and (4) a network generated using structural preference and homophily (S+H).

We chose to use the BA algorithm rather than similar network-generation algorithms as our baseline because it is more flexible in matching our ground-truth networks. For instance, the Watts-Strogatz model produces a fully connected network, which differs from our fractured ground-truth network. In contrast, the BA algorithm can generate a scale-free network with multiple separate components, similar to the visible Twitter network. Also, the BA algorithm can be easily parameterized to achieve a nearly identical structure, since the B model requires only the final node and edge count as parameters. In contrast, the Watts-Strogatz method requires the final node count, the average degree, and an alpha parameter that requires extensive tuning to generate realistic social networks.

Each of our networks is summarized in Table 2.

<table>
<thead>
<tr>
<th>Table 2: Basic network statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>GT</td>
</tr>
<tr>
<td>Node count</td>
</tr>
<tr>
<td>Edge count</td>
</tr>
</tbody>
</table>

4.1 Degree Distribution Comparisons

To begin, we compared the in-degree distribution of all networks in order to verify the expected scale-free nature of human social networks (Snijders, 2011). Removing all zero-degree nodes from each network, we arrive at the distribution of normalized degree frequencies in Figure 1.

![Figure 1: In-degree distribution](image-url)
However, the in-degree frequencies for networks generated by our models are consistently closer to the ground-truth frequencies than the BA network. This trend is quantified in Figure 2, which charts the relative deviation of each model’s degree distribution from the ground-truth distribution.

Note the consistently lower deviation of both S and S+H networks’ in-degree, particularly in the 1, 6, and 7-degree categories. The high deviations in the BA network imply that the GT network does not follow an ideal power-law distribution in the same way as the BA network. Lastly, all three models perform similarly in the 11-100 category, indicating that the GT network has a heavier tail than predicted. In nearly all other categories, the S and S+H networks achieve a lower deviation than the BA network.

We can further quantify this trend by calculating the mean deviations, as well as the Pearson correlation coefficient and Euclidean distance from the ground-truth distribution to the distributions of the three generated networks. We found the best-fitting model to be the one incorporating homophily, as it maximizes similarity and minimizes distance. The fitness statistics are summarized in Table 3, with the bold numbers highlighting the S+H network’s close fit with the GT network.

A similar trend favoring the homophily model was also found in the out-degree distributions of the four networks. Applying the same procedure to the networks’ out-degree distributions, we arrive at the deviations in Figure 3.

Note the consistently lower deviation of both S and S+H networks’ out-degree, particularly in the 1, 6, and 7-degree categories. The high deviations in the BA network imply that the GT network does not follow an ideal power-law distribution in the same way as the BA network. Lastly, all three models perform similarly in the 11-100 category, indicating that the GT network has a heavier tail than predicted. In nearly all other categories, the S and S+H networks achieve a lower deviation than the BA network.

Moving beyond low-level degree measures, we show that the homophily-dependent model yields emergent properties similar to ground-truth data. We selected three metrics to showcase the similarity of emergent properties: triadic closure, average path length, and giant component size. Triadic closure is calculated with the same formula outlined in Section 3, measured because of its correlation with short-range network cohesion. Average path length is equal to the mean of all shortest paths connecting nodes in the network, and we measured it because of its relation to clique-formation within a connected network. The average path length only includes paths between nodes within the giant component, which is the largest connected subgraph within the network and often includes the majority of nodes and edges. We also measured giant component size, because it provides a summary of large-
scale network cohesion which is key in predicting the breadth of information diffusion.

Table 4 outlines key structural metrics demonstrating the success of our model in replicating the GT network. The bold numbers indicate the closest fit to the GT statistics.

<table>
<thead>
<tr>
<th>Emergent property</th>
<th>GT</th>
<th>BA</th>
<th>S</th>
<th>S+H</th>
</tr>
</thead>
<tbody>
<tr>
<td>Triadic closure (undirected)</td>
<td>0.40</td>
<td>0.00</td>
<td>0.059</td>
<td><strong>0.075</strong></td>
</tr>
<tr>
<td>Average path length (undirected)</td>
<td>5.10</td>
<td>8.96</td>
<td><strong>6.31</strong></td>
<td>8.30</td>
</tr>
<tr>
<td>Giant component size (% of nodes included)</td>
<td>59.8%</td>
<td>91.9%</td>
<td>69.9%</td>
<td><strong>63.3%</strong></td>
</tr>
</tbody>
</table>

These numbers are more abstract to interpret but still demonstrate the success of our model in replicating ground-truth data. First, the GT network has an unexpectedly high rate of triadic closure, which is explained by the apparently high influence of mutual connections on link formation. Triadic closure was still matched best by the S+H model, indicating that homophily provides a boost when combined with our other structural metrics used in network generation.

Secondly, the average path length was best matched by the S network and was lower than anticipated in the GT network. This corroborates the implication that the GT network is denser, which yields a “small-world” effect. Still, the success of the S network suggests that our generation process outlined in Section 3 leads to a topology more realistic than the topology created by BA’s preferential attachment alone. The higher path length in the S+H network is likely related to homophilic clustering (explained later), since the average distance between nodes increases when the nodes exist in distinct clusters.

Thirdly, the giant component size was best matched by the S+H network and is a product of the network’s preferential attachment tendency. For instance, a model only incorporating preferential attachment (BA) tends to yield one core network rather than multiple networks (Newman 2002). This is clear when comparing the giant component size of the S and S+H networks, since the purely structural metrics that led to the S network overemphasized the importance of preferential attachment. Further, the addition of homophily in the S+H network dampened that effect and reduced the giant component to a more realistic size. Overall, these three emergent structural metrics demonstrate the success of our network generation process and particularly the importance of homophily in yielding high-level patterns.

In addition to purely structural metrics, we examined how homophily was correlated with apparent cluster formation in the ground-truth and synthetic networks. Visualizing the Twitter data in Figure 4 using Gephi (Bastian, Heymann and Jacomy 2009), we can infer a close connection between language and cluster structure. Node colors indicate user language and node size is relative to degree, such that nodes with more connections are larger.

We note particularly that non-English users tend to cluster by language. While connections between language clusters do exist, each language appears to cluster more often than not. A similar emergent cluster structure was replicated in our S+H network and can be seen in Figure 5, with identical node-coloring and node-sizing.
While the apparent grouping by homophily is an interesting observation, it does not guarantee that the language-clustering is identical in both networks. For instance, the GT language clusters appear to be denser than the S+H clusters. We verify this apparent pattern of language-clustering by partitioning the graph by language and calculating the density of each partition using

\[ D_G = \frac{|E_G|}{|V_G|(|V_G| - 1)} \]

where \( E_G \) and \( V_G \) represent the sets of edges and vertices in \( G \). We choose density as a metric for basic network coherence that scales predictably to networks of varying sizes. The densities are shown in Figure 6, which shows densities for the overall graph as well as each language’s subgraph (using abbreviations for the same languages as those displayed in Figures 4 and 5).

While not an exact fit, the correlation between the networks is noteworthy. For each language comparison, the S+H network achieves a similar density to the GT network, which can be seen in the relatively low absolute deviation between networks (\( \mu = 0.000829 \)). This is particularly evident in the French language cluster, which has the lowest relative density deviation between networks (13.4%). Based on density similarities, homophily by language is a clear factor in the GT clustered network structure and is emulated well by the S+H model.

Of course, it is important to remember that language is not the only demographic factor affecting edge formation. Future testing will help untangle the connections between partially-dependent demographics such as language and time zone. For instance, it is more likely for speakers of a language such as Ukrainian to represent fewer time zones than speakers of English, a more geographically diffuse language. Such overlaps between demographics could be addressed by a modified node-similarity function that incorporates conditional probabilities, such as \( P(\text{language} = \text{Ukrainian} | \text{time zone} = \text{UTC+2}) \).

5. Conclusions

In this paper, we described novel techniques for improving generative agent-based models of modern social media networks and demonstrated their effectiveness at improving synthetic network quality and realism by comparing produced networks to data about publicly available networks, in this case those found on the social media service Twitter. The continued improvement of this style of generative model, rooted in both network metrics and social science, is necessary for long-term understanding of the propagation of information, especially in partially-visible or wholly dark networks.

Future research in this area will extend and improve these models by incorporating additional social science-based agent-level parameters, including other traits not yet recognized. Further testing could also examine the ability of these models to correctly predict the structure of partially-visible or dark networks by examining the propagation of memetic information through partially-visible networks, such as Twitter, in comparison to the same kind of propagation in synthetic networks. Past studies (Peddinti, Ross, & Cappos, 2014) have proven that private traits across social media users are predictable, and this suggests a similar predictability among private users in general. Thus, future models will likely be able to model “dark” agents in social networks with little extra modification. This sort of testing will provide support for agent-based generative models as a means of uncovering the dynamics of network formation.
6. References


Author Biographies

**ELI STICKGOLD** is a scientist at Charles River Analytics Inc. whose research interests include machine learning techniques, cognitive models and network modeling.

**BRUCE SKARIN** is a consultant at Charles River Analytics Inc. whose research interests include system dynamics simulation, agent-based simulation, and network modeling.

**IAN STEWART** is a software engineer at Charles River Analytics Inc. whose research interests include network analysis, social media analysis, and noisy natural language processing.

**COREY LOFDAHL Ph.D.** is a senior scientist at Charles River Analytics Inc. whose research interests include modeling and simulation, quantitative analysis, and complex social systems.