Overview, Design concepts, and Details for the Online Social Network and Link Recommendation Model

The following is a description of the Online Social Network Model, presented using the Overview, Design Concepts, and Details format proposed by Grimm et al. (2006). Section 1 provides an overview of the model through discussion of its intended purpose, variables, and model generation process. This is followed by Section 2 which presents information on the model's stochastic processes, output, and emergent properties. Finally, Section 3 provides additional details regarding the model's initialization, inputs, and submodels. For those interested in the research conducted with this model, please see:

> **Sibley, C. and Crooks, A.T. (2020),** Exploring the Effects of Link Recommendations on Social Networks: An Agent-Based Modeling Approach, *Spring Simulation Conference (SpringSim'20),* Fairfax, VA.

1 Overview

1.1 Purpose

Recommender systems are omnipresent in online interactions, tailoring content to users by filtering through noisy data and identifying contextually relevant information. These personalized services inform the decisions and opinions of individuals, but it is unclear how these technologies may be impacting society. The purpose of this model is to explore how the most commonly used "friend-of-friend" or mutual connection-based link recommendations algorithms affect social network structure. This aims to begin addressing current debate about whether online recommendation systems fragment society into isolated echo chambers or expose individuals to more diverse communities and content (e.g., Hosanagar, Fleder, Lee, & Buja 2013; Sunstein 2018). This stylized model demonstrates how a simple agent-based model can be used to explore the effect of link recommendations on social network structure.

Agent-based modeling is an especially powerful methodology to utilize when real world data sets are challenging to acquire (e.g., real and/or complete social network data) (Crooks & Heppenstall

2012). In the case of recommender system research, many companies are willing to provide large data sets to assess prediction accuracy of various algorithms (e.g., Netflix Challenge (Bennett & Lanning 2007)), however social networking sites are generally unwilling to share data about users and/or their networks with the broader research community. As such, very little public research has explored the societal effects of recommendation systems. Stylized models with explicit assumptions, however, can lead to new understanding or knowledge about the complex system in question. Furthermore, the bottom-up nature of agent-based models allows investigation of how individual properties and behaviors can cause emergent patterns at the aggregate, or macro-level, such as power-law degree distributions in networks (Barabási & Albert 1999) or segregation within communities (Schelling 1971). In addition, conducting controlled experimentation within these artificial worlds enables exploration of different rules, assumptions, and causes for macro-level phenomena (Epstein & Axtell 1996).

Specifically, this model generates abstracted online social networks, by connecting nodes based upon varying proportions of a) connections from an underlying "real world" network (a small world, scale-free network) and b) link recommendations. Links that are formed by recommendation mimic the friend-of-friend algorithm, which is used on some of the largest social networking sites (e.g., Facebook, LinkedIn, Twitter). Essentially, these algorithms search for and recommend individuals with whom users share a large number of connections, but they themselves are not connected. Generated online networks are then analyzed to assess the influence that different proportions of link recommendations have on network properties, specifically: clustering, modularity, average path lengths, average eccentricity, network diameter, and degree distribution. For those less familiar with these social network analysis metrics, Table 1 provides brief definitions (and also density, which is relevant to the model generation). For more detailed explanations for these terms, readers are referred to Wasserman and Faust (1994) and Borgatti, Everett, and Johnson (2018).

1.2 State Variables and Scales

Given the purpose of this model is to explore network effects of link recommendations, the agents in this model represent individuals who are engaged on an online social network platform (e.g., Facebook, LinkedIn). Each agent is represented by a single node. Each node stores information about connections from the real world, i.e., connections/relationships that exist "offline." At each

time step, as an online network is continually generated as nodes probabilistically connect to 1) nodes from their real world network, and 2) nodes suggested by a link recommendation. Each node is assigned the same probability of connecting by each mechanism, for example, all nodes might have a 10% chance of connecting to a friend from the real world, and a 1% chance of connecting to a recommendation.

Table 1. Social Network Analysis definitions for metrics output by this model

Term	Brief Definition
Density	The percentage of the network that is directly connected (i.e, shares links), divided by the number of all possible connections. Values range 0 to 1.
Clustering coefficient	The proportion of a node's direct connections that are also directly connected to each other. Values range between 0 and 1.
Modularity	This analysis uses Louvain community detection method, and then measures the density of edges within communities divided by the density of edges outside communities. Values range between -1 and 1.
Path length	The shortest distance, or path, between two nodes, counted by the number of links. Also known as the geodesic. Values are greater than or equal to 1.
Eccentricity	The longest distance or path between two nodes without backtracking, counted by the number of links. Values are greater than or equal to 1.
Diameter	The shortest path length between the two most distant nodes in the network. Values are greater than or equal to 1.
Degree	The number of nodes that a node is linked to, not counting itself. Values are all greater than or equal to 0.

The underlying real world network is formed at the beginning of each simulation, using the power-law cluster graph (Holme & Kim 2002) generation algorithm within NetworkX (Hagberg, Swart, & Chult 2008). This specific graph generation is used to initialize a network exhibiting properties characteristic of real world social networks: small average geodesic, high clustering coefficient, and scale-free degree distribution (Wang & Chen 2003). A 10,000 node example real world network is depicted in Figure 1. This graph generation algorithm is similar to preferential attachment rules (Barabási & Albert 1999) but adds a triangle with some probability after each edge is created, in order to promote high levels of clustering, which are absent in the original

preferential attachment model. The following parameter settings are used within NetworkX's "powerlaw_cluster_graph" generator, in order to mirror Holme and Kim's (2002) output: n = 10,000 (number of nodes); m = 3 (number of edges added to each node, at each step); and p = 0.5 (probability of adding a triangle following the addition of a random edge, at each step), but different graph generators and/or size settings can also be explored, depending on the underlying "real world" network features of interest.

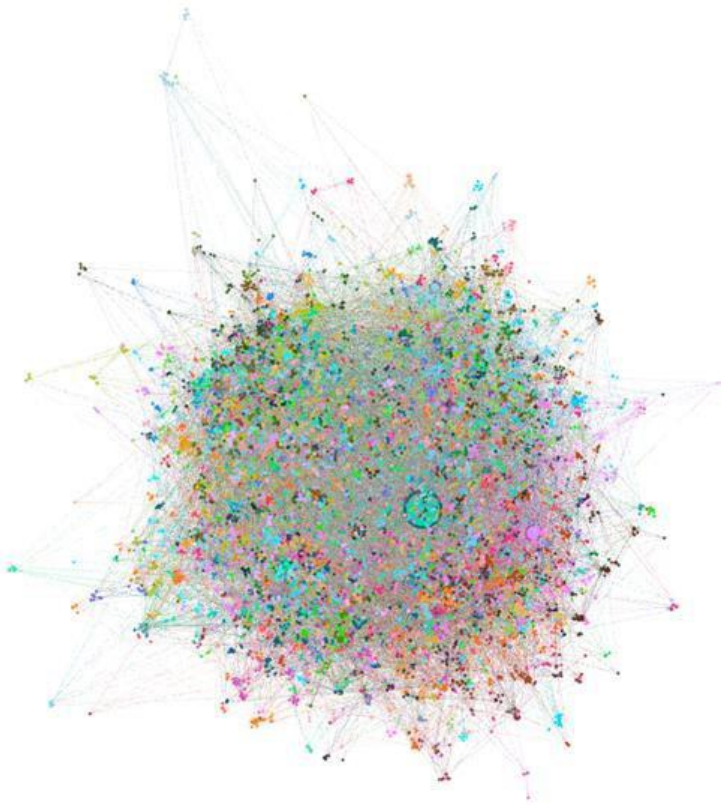

Figure 1. Real World network with 10,000 nodes colored by modularity class and sized by degree

The heterogeneity among nodes exists in their a) real world ego-networks and b) mutual shared connections with non-connections. Individual nodes store a list of their connections from the real world network, which represent their 'friends' from the real world. This information is relevant in the formation of an online network, since individuals will be likely to connect online with people who they know from the real world. Individual nodes do not have any attributes, such as demographic information, since this model's purpose is simply to see how network structure is impacted by link recommendations. However, nodes could easily be initialized and calibrated with attributes in the event follow-on work may want to explore how various attributes disperse

in an online network. Furthermore, space is not spatially explicit in this model. Links among nodes represent friendship ties and do not include any environmental or geographical information.

1.3 Process Overview and Scheduling

The most important process within this model is the online network generation, which takes place in two steps, shown in Table 2, through an example network of 10,000 nodes. As shown in Step 1 of Table 2, each simulation instance begins with the creation of a real world network and calculation of its network density, which is used as a stop function for when to finish adding connections within the online social network. Holding network density constant across all real world and online networks was chosen to ensure that density differences wouldn't contribute to any differences in metrics of interest (e.g., clustering, average path length), and instead the effects could be attributed solely to the link recommendations. This setting, however, could be modified if further exploration desired the online network to grow to a lessor or greater size (by simply multiplying the density 'stop function' value by some constant, such as 0.9 or 1.1). At the end of this step, each node stores a list of its connections, which will be referenced in Step 2.

In Step 2 of Table 2, the online social network is generated by iteratively making connections among the 10,000 nodes, using a combination of real world friendships and link recommendations. To further expound on the link recommendation algorithm - a mutual connection is defined as an individual (i.e., node) with whom a node is not directly connected, however they share a common connection. The simplest example of this is an open triad, where A and B are connected, and A and C are connected, but B and C are not connected. Here, B and C share A as a mutual connection and so it is predicted that they are more likely to like or know one another, as compared to another node with whom no mutual connections exist. As such, the link recommendation algorithm counts the number of shared connections for every pair of nodes that are not already connected. For each individual node it then identifies the node with whom it shares the greatest number of mutual connections. This node is suggested for recommendation. If there are multiple nodes that share the same maximum number of mutual connections, then one is chosen at random so no preference is given to specific node indexes.

Table 2. Online Social Network generation process

Step 1	**Generate Real World Network (Holme Kim model: n = 10,000, m =3, p = 0.5) and store network density**
Step 2	Generate Online Social Network, starting with 10,000 isolates
2a	Each node links with a connections from Real World Network, with p=0.10, set by the experimenter
2b	Update mutual connection count for each pair of nodes that are not already connected
2c	Each node links with whomever they share the maximum number of mutual connections but are not currently connected, with p=0.01, set by the experimenter
2d	Repeat Steps 2a - 2c until Online Social Network density > Real World density

As shown in Step 2a (of Table 2) the online social network starts as a group of 10,000 isolates (or whatever number of nodes is set for the real world network). At each new time step, every node has some probability (set by the user, but 0.10 was chosen for this example) that it will form a link with one of its real world connections, to whom it is not already connected. This connection is chosen at random from a list of available options. After all nodes have completed Step 2a every node updates: 1) their list of connections, and 2) stores the count of mutual connections they have with all non-connections. Synchronous updating at this stage (Step 2b) of online network formation, and not within the next step, was chosen so that the system was not biased towards suggesting nodes that become more connected within a round of updating. In other words, this was done so that there was no advantage to some nodes simply due to them forming their connections first.

In Step 2c, information about the number of mutual connections, shared with non-connections, is then utilized and each node, with some probability (set by the user, but 0.01 for this example), is given the opportunity to form an edge based upon the link recommendation. This process of

forming connections within the online social network continues until the online network's density equals that of the real world network. Additionally, the connection probabilities that are selected (via real world or link recommendation) influence the number of iterations before the network is fully formed and reaches maximum density: with higher probabilities (e.g., 40%) resulting in fewer rounds and lower probabilities (e.g., 5%) resulting in more rounds of development and greater asynchrony (i.e., fewer individuals making connections within each round).

2 Design Concepts

2.1 Stochasticity

Stochasticity is seen when each node forms connections during each round (Steps 2a and 2c), was chosen to simulate the asynchronous nature of online social networks growing over time and only some portion of users logging in and connecting to new individuals, at each time step. Furthermore, it is assumed that individuals participating on online social networks would have at least the same or a greater proportion of connections with individuals who they actually know in the real world, rather than shared contacts (Ellison, Steinfield, & Lampe 2007). As such, it is suggested that the probability of connecting via link recommendation should be set with the same or a lower probability than connecting to a real world contact. This assumption holds for sites such as Facebook and LinkedIn, where links are reciprocal and interactions tend to be more intimate than, for example, Twitter, which has directional links and serves as both entertainment (following celebrities) and socializing (Quan-Haase & Young 2010; Johnston, Tanner, Lalla, & Kawalski 2013).

2.2 Observation

The model will output both the real world network and the online social network in GraphML format, which can then be loaded and analyzed in any software package of choice. Additionally, the model will output a .csv file of both the real world and online social network degree counts per node which can be used for quick observation and analysis of the degree distributions. The source code also has numerous print statements that allows for testing and verifying that the code is behaving as intended. Additionally, the social network analysis script (SNAMetricAnalysisOpenABM.py) will output the following network metrics: number of nodes; average clustering coefficient, modularity, average eccentricity, diameter, degree of fragmentation,

average degree, maximum degree, and average geodesic. Finally, the power-law analysis script (Power+law+analysis.py) will provide power-law fit line estimates and statistics, and output graphs of the degree distributions, comparing lognormal, exponential, and power-law fits.

2.3 Emergence

The network-level metrics and degree distribution emerge from the individual behaviors of the nodes. The real world network imposes a certain order on the nodes that would result in a clustered, power-law distribution if the link recommendations were set to zero. However, once the additional link recommendations are added, the network structure and degree distribution is altered due to the individual nodes adapting to their changing environment, i.e., ego-network and network of once-removed connections.

3 Details

3.1 Initialization

At the beginning of each simulation instance, a new underlying real world network is formed using the power-law cluster graph (Holme & Kim 2002) generation algorithm within NetworkX (Hagberg, Swart, & Chult 2008), with the following parameter settings: n = 10,000 (number of nodes); m = 3 (number of edges added to each node, at each step); and p = 0.5 (probability of adding a triangle following the addition of a random edge, at each step). Each node stores its connections from the real world, which are used in Step 2a to form connections online. These values were chosen to mirror the results presented in Holme and Kim's (2002) analysis, however, a 10,000 node network was used here instead of their 100,000 node network. Different graph generators within NetworkX can easily be swapped in and explored, however, depending on the experiment being performed. The other initialization settings are the probabilities of linking online with a real world connection (Step 2a) and a recommendation-based connection (Step 2c). The real world setting of 10% was chosen ultimately with the purpose of adding asynchrony to the formation of the online network; such that not every node formed a connection every round. Values equal to and below 10% were selected for the recommendation probability (Step 2c) based on the assumption that people would have equal or fewer numbers of recommendation-based links.

3.2 Input

As described in Section 3.1, there are 3 input values that are used to initialize the network:

1) The underlying real world model (e.g., 10,000 node power-law cluster network)

2) The probability of connecting based upon a real world contact (e.g., 10%)

3) The probability of connecting based upon a link recommendation (e.g., 1%)

3.3 Submodels

As mentioned above, the real world model that initializes each node's real world connections uses Holme and Kim (2002) method of creating a graph that is scale-free and has high clustering. The link-recommendation algorithm, also described above, identifies every node on the online network to whom a node is not already connected (i.e., non-connections), and then counts the number of connections that it has in common with each of those non-connections. For each node, the algorithm then connects that node to the non-connection with whom it shared the most connections. If multiple nodes have the same maximum value, then a node is chosen at random. This algorithm is based off publically available descriptions of friend-of-friend link recommendations used by large social network sites (Facebook 2019; LinkedIn 2019).

4 References

Barabási, A.L., and R. Albert. 1999. "Emergence of Scaling in Random Networks". *Science* vol. 286, pp. 509-512.

Bennett, J., and S. Lanning. 2007. "The Netflix Prize". In *Proceedings of the KDD Cup Workshop*, pp. 3-6. New York, ACM.

Borgatti, S. P., M.G. Everett, and J. C. Johnson. 2018. Analyzing social networks. Sage.

Crooks, A. T., and A. J. Heppenstall. 2012. Introduction to Agent-based Modelling. In *Agent-based Models of Geographical Systems*, pp. 85-105. Dordrecht, Springer.

Epstein, J. M., and R. Axtell. 1996. *Growing Artificial Societies: Social Science from the Bottom Up*. Santa Fe, NM, MIT Press.

Facebook. 2019. "Finding Friends and People You May Know". https://www.facebook.com/help/findingfriends. Accessed April 20, 2017.

Grimm, V., U. Berger, F. Bastiansen, S. Eliassen, V. Ginot, J. Giske, J. Gross-Custard et al.. 2006. "A Standard Protocol for Describing Individual-based and Agent-based Models". *Ecological Modelling* vol. 198, pp. 115-126.

Hagberg, A., Swart, P., and D. Chult. 2008. “Exploring Network structure, dynamics, and function using NetworkX”. In *Proceedings of the 7th Python in Science Conference* (SciPy2008), pp. 11–15.

Holme, P., and B. J. Kim. 2002. “Growing Scale-free Networks with Tunable Clustering”. *Physical Review E* vol. 65, pp. 1-4.

Hosanagar, K., D. Fleder, D. Lee, and A. Buja. 2013. “Will the Global Village Fracture into Tribes? Recommender Systems and their Effects on Consumer Fragmentation”. *Management Science* vol. 60, pp. 805-823.

Johnston, K., M. Tanner, N. Lalla, and D. Kawalski. 2013. "Social Capital: The Benefit of Facebook ‘friends’." *Behaviour & Information Technology* vol. 32, pp. 24-36.

LinkedIn. 2019. “People You May Know Feature – Overview”. https://tinyurl.com/uvyt2ts. Accessed April 20, 2017.

Quan-Haase, A., and A. L. Young. 2010. "Uses and Gratifications of Social Media: A Comparison of Facebook and Instant Messaging." *Bulletin of Science, Technology & Society* vol. 30, pp. 350-361.

Schelling, T. C. 1971. “Dynamic Models of Segregation”. *Journal of Mathematical Sociology* vol. 1, pp. 143-186.

Sunstein, C. R. 2018. *# Republic: Divided Democracy in the Age of Social Media*. Princeton, NJ, Princeton University Press.

Wang, X. F., and G. Chen. 2003. “Complex Networks: Small-world, Scale-free and Beyond”. *IEEE Circuits and Systems Magazine* vol. 3, pp. 6-20.

Wasserman, S., and K. Faust. 1994. Social network analysis: Methods and applications. Vol. 8. Cambridge university press.