Growth mechanisms in continuously-observed networks: Communication in a Facebook-like community

Tore Opsahl¹ and Bernie Hogan²

¹Imperial College Business School, Imperial College London; ²Oxford Internet Institute, University of Oxford

Method for analyzing tie generation

Motivation
1. Although more and more data is collected in real-time or contain exact time information, few methods exist to analyze these data
2. Static data is an issue for methods analyzing tie growth

Existing methods and frameworks
1. Simple descriptive measures
   • What: Clustering coefficient measures the level of clustering in the network. If an above random coefficient is found, then it is assumed a triadic closure effect is at work (e.g., Opsahl and Panzarasa, 2009)
   • Why not: Clustering might also be a result of homophily, and without using a multivariate model, the effects cannot be distinguished
2. Exponential Random Graph Models (ERGM; Robins and Morris, 2007)
   • What: Predicts substructures
   • Why not: Hard to define normalizing constant; only small networks
3. Simulating the evolution of ties (SIENA; Snijders, 2001)
   • What: When multiple snapshots of the network is known, methods have been developed for simulating the network evolution
   • Why not: Simulations might distort the actual effects; only small networks
4. Survival / hazard models (Butts, 2008; de Nooy, 2010)
   • What: Predicts the time to tie creation
   • Why not: Context specific; only small networks

Proposed framework

Conditional logistic regression

1. At a given time t, a node j decides to form a tie.
2. This tie can be directed towards the set of available nodes in the network at that time, A(t). If analyzing a binary network, this set would be assumed to include all the nodes in the network at time t, or j is currently not tied to. Conversely, if analyzing a network in which nodes can form multiple ties between them, then A(t), would include all the nodes in the network at time t.
3. The node that receives the tie, node i, can have a number of properties, Zj,t. The purpose of the conditional logistic regression model is to see whether the properties of node i, Zj,t, stand out from the properties of all the available nodes, Zj,t. The model can be formalized as:

\[
P(Y_{ij} = 1 | Z_{j,t}) = \frac{exp\left(\beta^T Z_{j,t}\right)}{1 + exp\left(\beta^T Z_{j,t}\right)}
\]

4. This model directly probes the decision-making process of nodes. By combining LG, this model can be scaled-up to analyze millions of nodes.

References
de Nooy, W., 2010. Networks of action and events over time. Social Networks, in press.

Communication in a Facebook-like community

Data (Opsahl and Panzarasa, 2009)
1. An online profile-based community oriented towards college students at University of California-Irvine in 2004.
2. The profiles contained user-inputted fields (such as age, gender, and school affiliation) and user-generated content (such as blog postings and forum discussions) as well as automated fields such as the number of times a profile was visited and a list of a user’s declared friends.
3. The site enabled individuals to search for others based on keywords. Thus, it enabled several social affordances that should influence the resulting topology, such as direct messaging and replies, counts of friends, and demographic information. It lacked the ability to have comments on one’s profile page as well as multi-recipient messaging.
4. The dataset is a cleaned version of the private messages or dyadic exchanges on the site. In short, it is a set of 1,899 nodes that collectively sent 59,835 messages over 20,296 directed ties among them. The dataset only included the time of sending and anonymized node identification numbers for the sending and receiving users. The content of messages was not available.

Model specification

1. As the aim of this analysis is to determine the drivers behind communication, we defined the panels in the model, t, to be each message sent across the site. For each t, the real receiver of the message is known, and 19 control nodes are randomly picked among users registered on the site at that time.
2. Tie predictors
   • Homophily (similar age, same marital status, and same gender)
   • Focus constraints (same place of origin, same school, and similar year of study)
   • Reinforcement (above random reinforcement)
   • Reciprocity (above random reciprocity)
   • Triadic closure (above random clustering)
   • Preferential attachment (power-law-like in-degree distribution)

Findings

1. Homophily and focus constraint matter
2. Gender heterophily (e.g., due to dating)
3. Past interaction increase likelihood of communication 14 times
4. Users are 23 times more likely to reply
5. Triadic closure is a weak effect
6. Each communication partner increase a likelihood of receiving a tie by 1% only (weak effect of preferential attachment)

Implications and Areas of applicability

1. Understanding communication patterns in large-scale datasets
2. Understanding users’ decisions
3. Determining probabilities of communication
4. Predicting future communication
5. Calibrating algorithmic recommender systems

Method for analyzing tie generation

Motivation
1. Although more and more data is collected in real-time or contain exact time information, few methods exist to analyze these data
2. Static data is an issue for methods analyzing tie growth

Existing methods and frameworks
1. Simple descriptive measures
   • What: Clustering coefficient measures the level of clustering in the network. If an above random coefficient is found, then it is assumed a triadic closure effect is at work (e.g., Opsahl and Panzarasa, 2009)
   • Why not: Clustering might also be a result of homophily, and without using a multivariate model, the effects cannot be distinguished
2. Exponential Random Graph Models (ERGM; Robins and Morris, 2007)
   • What: Predicts substructures
   • Why not: Hard to define normalizing constant; only small networks
3. Simulating the evolution of ties (SIENA; Snijders, 2001)
   • What: When multiple snapshots of the network is known, methods have been developed for simulating the network evolution
   • Why not: Simulations might distort the actual effects; only small networks
4. Survival / hazard models (Butts, 2008; de Nooy, 2010)
   • What: Predicts the time to tie creation
   • Why not: Context specific; only small networks

Proposed framework

Conditional logistic regression

1. At a given time t, a node j decides to form a tie.
2. This tie can be directed towards the set of available nodes in the network at that time, A(t). If analyzing a binary network, this set would be assumed to include all the nodes in the network at time t, or j is currently not tied to. Conversely, if analyzing a network in which nodes can form multiple ties between them, then A(t), would include all the nodes in the network at time t.
3. The node that receives the tie, node i, can have a number of properties, Zj,t. The purpose of the conditional logistic regression model is to see whether the properties of node i, Zj,t, stand out from the properties of all the available nodes, Zj,t. The model can be formalized as:

\[
P(Y_{ij} = 1 | Z_{j,t}) = \frac{exp\left(\beta^T Z_{j,t}\right)}{1 + exp\left(\beta^T Z_{j,t}\right)}
\]

4. This model directly probes the decision-making process of nodes. By combining LG, this model can be scaled-up to analyze millions of nodes.

References
de Nooy, W., 2010. Networks of action and events over time. Social Networks, in press.