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

Tore Opsahl[1] and Bernie Hogan[2]

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

Most network studies rely on static data, which creates methodological issues when predicting tie creation. Although there has been a surge in continuously-observed datasets (i.e., non-static), few methods exist to study these. This poster proposes a framework for assessing multiple growth mechanisms (e.g., homophily, focus constraints, reinforcement, reciprocity, triadic closure, and preferential attachment) in such datasets, and applies it to communication within a Facebook-like community. While some mechanisms are significant independently, they are insignificant in multivariate analyses. This finding exemplifies that descriptive measures, such as the clustering coefficient, cannot be relied upon for studying mechanisms of tie generation.

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 especially 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 assume that 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 *i* 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* that node *i* 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 *j*, can have a number of properties, $Z_{j,t-1}$. The purpose of the conditional logistic regression model is to see whether the properties of node *j*, $Z_{j,t-1}$, stand out from the properties of all the available nodes, $Z_{At,t-1}$. The model can be formalized as:

$$P\{j_t = j|Z_{t-1}\} = \frac{\exp(\beta' Z_{j,t-1})}{\sum_{h \in At} \exp(\beta' Z_{h,t-1})}$$

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

References

Butts, C.T., 2008. A relational event framework for social action. Sociological Methodology, 38: 155–200.

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

Opsahl, T., Panzarasa, P., 2009. Clustering in weighted networks. Social Networks. 31 (2), 155-163.

Robins, G.L., Morris, M., 2007. Advances in exponential random graph (*p**) models. Social Networks. 29 (2), 169-172.

Snijders, T.A.B., 2001. The statistical evaluation of social network dynamics. Sociological Methodology. 31, 361-395.

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. The profiles contained user-inputed 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.
2. 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.
3. 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 the 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

	Univariate		Multivariate	
Model	1-15		16	
Similar age	0.7000 *	101%	-0.0605	-6%
	(0.2837 ; 6*)		(0.1016)	
Same marital status	0.3451 ***	41%	0.2652 ***	30%
	(0.0343 ; 101***)		(0.0261)	
Same gender	-1.5430 ***	-79%	-1.0329 ***	-64%
	(0.0592 ; 680***)		(0.0462)	
Similar area of origin	0.4723 ***	60%	0.3968 ***	49%
	(0.0417 ; 128***)		(0.0278)	
Same school	0.2431 ***	28%	0.1988 ***	22%
	(0.0491 ; 24***)		(0.0328)	
Similar year of study	1.3484 ***	285%	0.9382 ***	156%
	(0.1058 ; 162***)		(0.0877)	
Reinforcement (0/1)	4.5661 ***	9517%	2.6873 ***	1369%
	(0.0845 ; 2920)		(0.0879)	
Reinforcement	0.7598 ***	114%		
	(0.0759 ; 100)			
Reciprocity (0/1)	5.0482 ***	15474%	3.1960 ***	2343%
	(0.0823 ; 3764***)		(0.0910)	
Reciprocity	1.0608 ***	189%		
	(0.0804 ; 174***)			
Triadic closure	0.3450 ***	41%	0.0881 ***	9%
	(0.0398 ; 75***)		(0.0211)	
In-degree	0.0433 ***	4%	0.0142 ***	1%
	(0.0007 ; 3855***)		(0.0010)	
In-strength	0.0110 ***	1%		
	(0.0002 ; 3517***)			
Wald X^2			3,221 ***	

Growth mechanisms tested in a conditional logistic regression framework with 19 control cases for each observed case. To ensure comparability across models, the same set of control nodes was used. Robust standard errors adjusted for clusters based on sender are in parentheses. For univariate analyses, the standard errors are followed by Wald X2 scores. N=1,196,400 (59,820 strata x 20 observations).
† p < 0.10; * p < 0.05; ** p < 0.01; *** p < 0.001.

read the full paper: arXiv:1010.2141 – try the framework: *tnet* from version 3.0