

Identification of Churn and Fraud Communities in Large-Scale Customer Networks

Torsten Dierkes, Martin Bichler, and Ramayya Krishnan*
Department of Informatics, TU München, Germany
*Carnegie Mellon University, USA
{torsten.dierkes, bichler}@in.tum.de

Summary

Community structure is a feature of complex networks (Girvan & Newman, 2002). Although the topology of complex networks does not follow obvious patterns, other unexpected statistical features have been identified that are surprisingly common to diverse networks. The “small world” feature predicts that the average distance between two arbitrary nodes is surprisingly short (Milgram, 1967). Scalefreeness predicts that the probability of a single node having a certain number of contacts in the network does not depend on the network’s size (Barabási & Albert, 1999). The feature we focus on in this work is that nodes of a network often belong to groups or communities that are only indirectly observed by having many within-group connections and only few connections to other groups. The term community structure was coined for this observation.

Communities can be of interest to marketers. It might be interesting to target a community of “highly connected” customers in a marketing campaign for a new product. Communities of customers, who are likely to churn are important for win-back activities. It would also be interesting to find communities of customers with fraud behavior. In our research, we explore techniques to identify communities of customers with a high churn or fraud likelihood.

For this purpose, we analyze data from a European telecommunications provider with several million customers. We conjecture that the *communication between these people will exercise influence* and that *community members will exhibit some level of similarity*. The tendency of individuals to associate with others similar to themselves is termed homophily (Lazarsfeld & Merton, 1954). Indeed, homophily is documented as one of the most robust and pervasive features of social networks across a wide array of characteristics, including age, race, profession, and patterns of behavior (see McPherson, Smith-Lovin & Cook, 2001 for a survey). For example, Christakis & Fowler (2007) found that individuals linked with obese friends have an up to 57% increased probability of becoming obese themselves.

The methods for detecting communities are relatively new. Most algorithms have been developed recently — within the last few years. Although many publications share the term ‘community’, the definition varies. The strictest one is that of a clique. While this definition is straightforward, finding cliques is not – the fundamental problem of finding cliques is NP-hard and thus is intractable for the present large-scale networks. Other explicit community definitions in literature are based on k-cliques (Palla, Derényi, Farkas & Vicsek, 2005), edge connectivity (Hartuv & Shamir, 2000), or highly connected subgraphs (Radicchi, Castellano, Cecconi, Loreto & Parisi, 2004).

Some approaches are based on well-founded statistical concepts of communities, e.g. stochastic blockmodeling estimates the posterior probability distribution of all possible community structures based on the respective statistical model of a community (Nowicki & Snijders, 2001). While statistical community identification techniques allow for well-founded and application-specific community modeling, they usually suffer from low scalability, which is the reason why they are not the appropriate choice for large-scale networks.

Many classes of community identification algorithms employ implicit community definitions implied by their underlying algorithmic approach, e.g. repeated bisection or edge removal techniques which repeatedly remove network edges or bisect the network, respectively, or hierarchical clustering which adds new nodes to groups successively by examining how similar pairs of nodes are.

More recently, fast heuristics have been developed to address scalability issues with large-scale networks. Raghavan, Albert & Kumara (2007) developed a simple label propagation algorithm in which densely connected groups of nodes form a consensus on a unique label to form communities, yielding a near linear time complexity. Clauset, Newman & Moore (2004) developed a hierarchical agglomeration algorithm (CNM) with time complexity $O(m \log n)$ which was further improved by Wakita & Tsurumi (2007) by a heuristic that attempts to merge community structures in a balanced manner (CNM2), improving time complexity by orders of magnitude. Minimum Spanning Trees (MSTs) (Prim, 1957; Kruskal, 1956) use edge removal techniques based on well known MST algorithms, yielding a time complexity of $O(m \log n)$. Danon, Diaz-Guilera, Duch & Arenas (2005) compared some of the community finding algorithms.

While these heuristics generate community structure for large-scale networks in practically feasible time frames, it remains unclear, whether the identified communities provide valuable information for marketers, since the communities are based on the heuristic's implicit community definition and not the marketers' ones.

We apply community identification algorithms to a large-scale customer network in the telecommunication industry and analyze the characteristics of resulting networks. The algorithms are applied *to the entire network*, but also *to a network of pre-selected nodes* that satisfy certain criteria (e.g., only customers, who have cancelled their contract). We will provide descriptive statistics and classification models identifying characteristics and homophily properties of respective customer communities and network topology measures such as the density or cohesion within a cluster.

The contribution of this paper is as follows: This is the first paper to apply a set of new and scalable community identification algorithms to a large-scale customer network of a telecommunication provider. Much work on community identification and graph clustering algorithms is motivated by respective applications (Brandes & Erlebach, 2005). We show, however, that the resulting clusters exhibit little information from a marketing point of view when applied to customer networks generated from call detail records. In contrast, a simple heuristic, first selecting churn customers, and then analyzing their communication traits, can reveal much useful information for a marketer, and help them in designing win-back campaigns, which is typically what marketers aim for.

Reference List

- Barabási, A.-L., & Albert, R. (1999). Emergence of Scaling in Random Networks. *Science*, 286(5439), 509–512
- Brandes, U., & Erlebach, T. (2005). *Network analysis: Methodological foundations*, from <http://www.springerlink.com/openurl.asp?genre=issue&issn=0302-9743&volume=3418>
- Christakis, N. A., & Fowler, J. H. (2007). The Spread of Obesity in a Large Social Network over 32 Years. *The New England Journal of Medicine*, 357(4), 370–379
- Clauset, A., Newman, M. E. J., & Moore, C. (2004). Finding community structure in very large networks. *Phys. Rev. E*, 70(6), 66111
- Danon, L., Diaz-Guilera, A., Duch, J., & Arenas, A. (2005). Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, (9).
- Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(7821-7826).
- Hartuv, E., & Shamir, R. (2000). A clustering algorithm based on graph connectivity. *Information Processing Letters*, 76(4-6), 175-181, from <http://citeseer.ist.psu.edu/hartuv99clustering.html>
- Kruskal, J. B. (1956). On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem. *Proceedings of the American Mathematical Society*, 7(1), 48–50
- Lazarsfeld, P. F., & Merton, R. K. (1954). Friendship as a social process: A substantive and methodological analysis. In M. Berger, T. Abel & C. Page (Eds.), *Freedom and Control in Modern Society* (pp. 18-66). New York: Van Nostrand.
- McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in Social Networks. *Annual Review of Sociology*, 27, 415–444. Retrieved September 18, 2008
- Milgram, S. (1967). The Small World Problem. *Psychology Today*, 2, 60–67
- Nowicki, K., & Snijders, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96, 1077–1087
- Palla, G., Derényi, I., Farkas, I., & Vicsek, T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043), 814-818, from <http://dx.doi.org/10.1038/nature03607>
- Prim, R. C. (1957). Shortest connection networks and some generalizations. *Bell System Technology Journal*, 36, 1389-1401
- Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., & Parisi, D. (2004). Defining and identifying communities in networks. *Proc Natl Acad Sci U S A*, 101(9), 2658-2663
- Raghavan, U. N., Albert, R., & Kumara, S. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 76(3).
- Wakita, K., & Tsurumi, T. (2007). Finding Community Structure in Mega-scale Social Networks. *CoRR*,