

Networks, Communities and Kronecker Products

Jure Leskovec
Stanford University
jure@cs.stanford.edu

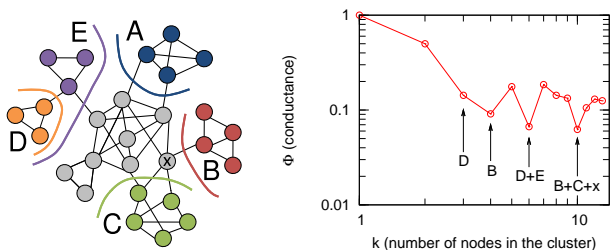


Figure 1: An example network and a corresponding Network Community Profile (NCP) plot.

ABSTRACT

Emergence of the web and online computing applications gave rise to rich large scale social activity data. One of the principal challenges then is to build models and understanding of the structure of such large social and information networks. Here I present our work on clustering and community structure in large networks, where clusters are thought of as sets of nodes that are better connected internally than to the rest of the network. We find that large networks have very different clustering structure from well studied small social networks and graphs that are well-embeddable in a low-dimensional structure. In networks of millions of nodes tight clusters exist at only very small size scales up to around 100 nodes, while at large size scales networks becomes expander like. A network model based on Kronecker products efficiently models such core-periphery network structures. The results suggest broader implications for data analysis and machine learning in sparse and noisy high-dimensional social and information networks, where intuitive notions about cluster quality fail.

Categories and Subject Descriptors: H.2.8 Database Management: Database applications – Data mining

General Terms: Measurement; Experimentation.

Keywords: Social networks; Graph partitioning; Community structure; Conductance; Network community profile; Kronecker graphs.

1. INTRODUCTION

Network communities, usually thought of as groups of nodes with better connections between its members than with the remainder of the network, represent basic structures for understanding the organization of complex networks [4, 9, 3]. While community structure in small networks has been well established, the question of how it scales to large networks has remained largely unanswered.

Copyright is held by the author/owner(s).
CNMKM'09, November 6, 2009, Hong Kong, China.
ACM 978-1-60558-807-0/09/11.

2. NETWORK COMMUNITY PROFILE

We model each network by an undirected graph, in which nodes represent entities and edges represent interactions. In order to formalize the notion of a community, we adopt the notion of *conductance* $\phi(S)$ of a set of nodes S , which may be thought of as the ratio between the number of connections pointing outside the cluster S and the number of connections inside S . More formally, *conductance* $\phi(S)$ of a set of nodes S is $\phi(S) = |\partial S|/\text{Vol}(S)$, where $|\partial S|$ denotes the size of the edge boundary, $|\partial S| = |\{(u, v) : u \in S, v \notin S\}|$, and $\text{Vol}(S) = \sum_{u \in S} d_u$, where d_u is the degree of node u . For example, in Fig. 1 since $\phi(A) = \frac{2}{14} > \phi(B) = \frac{1}{11}$, the set of nodes B is more community-like than the set A .

Using conductance as a measure of network community quality, we then define the *Network Community Profile (NCP)* which plots the conductance score of the best possible cluster on k nodes, $\Phi(k) = \min_{|S|=k} \phi(S)$, as a function of k (Fig. 1). For example, for $k = 4$, among all sets of 4-node clusters, B has best (*i.e.*, minimum) conductance, and thus $\Phi(4) = \frac{1}{11}$. Similarly, D and $D+E$ denote the best conductance sets on 3 and 6 nodes, respectively. Just as the magnitude of the conductance provides information about how community-like is a set of nodes, the shape of the NCP provides insight into how well expressed are network communities as a function of their size. Although the NCP is intractable to compute exactly, approximation algorithms for graph partitioning can be used to approximate it [1].

The NCP behaves in a characteristic downward-sloping manner (Fig. 2(a) and [8, 7]) for mesh-like graphs, *e.g.*, low-dimensional lattices, road networks, and random geometric graphs. The NCP is also downward-sloping for networks with hierarchical community structure [2] and for small commonly-studied social networks.

However, the community structure in large networks (Fig. 2(b, c)) differs substantially from the structure found in small networks [8].

First, up to a size scale of roughly 100 nodes, the NCP is downward-sloping, which means that as community size increases the best possible communities get progressively better (Fig. 2(b, c)). Second, the global minimum of the NCP occurs at a size scale of roughly 100 nodes which is practically independent of the network size, suggesting that there is a natural size scale to network communities. Third, above this size scale, the NCP gradually increases. The upward slope of NCP suggests, and empirically we observe, that as a function of increasing size, the communities become more and more attached to the remainder of the network. That is, as community size increases the number of edges pointing outside the community grows relatively faster than the number of edges inside. Thus, at larger size scales, even the existence of communities as sets of nodes with better internal connectivity than external connectivity lacks empirical support (Fig. 2(b, c)). Fourth, even at the largest size scales, these real networks have significantly more structure than random networks with same degree sequence.

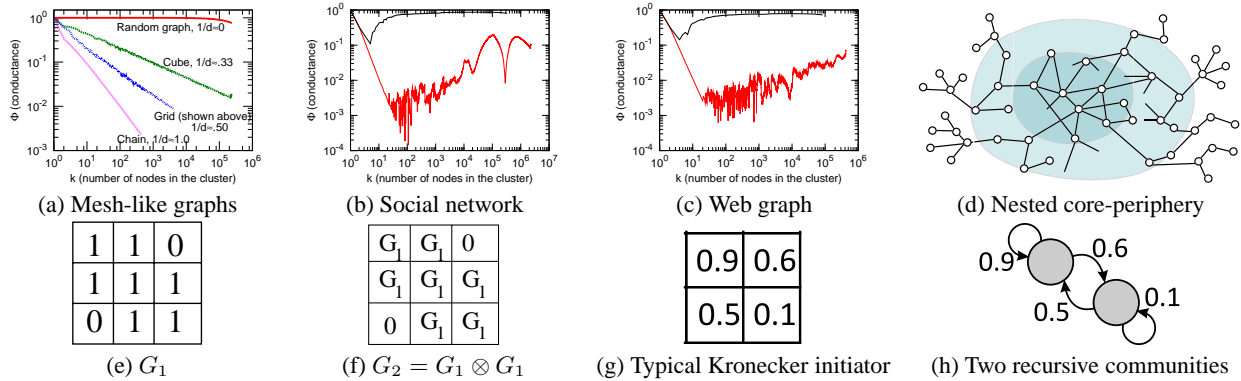


Figure 2: NCPs for mesh-like networks and small social networks slopes downwards (a). For large networks, like friendships in LiveJournal (b) and a Google web graph (c), it slopes downward and then upwards, which suggests a onion like core-periphery structure (d) where the network gets denser and denser as we move towards the center of the network. (e, f) Kronecker multiplication. (g, h) Kronecker initiator matrix can be thought of as two communities, where each sub-community can then be recursively divided.

3. KRONECKER GRAPHS

Such NCP can be explained by a “core-periphery” structure, in which the network consists of a large moderately well-connected core (upward part of NCP) and a large number of small communities barely connected to the core (downward part of NCP) (Fig. 2(d)). Kronecker graphs [5, 6] are a natural model for such network structure. Figure 2(e,f) shows the recursive construction of Kronecker graphs. We start with adjacency matrix G_1 , and Kronecker power it to obtain a larger graph G_2 . To produce G_k from G_{k-1} , we “expand” nodes of G_{k-1} by copies of G_1 , and join the copies according to the adjacencies in G_{k-1} . Intuitively, communities in the graph grow recursively, with communities getting recursively expanded into miniature copies of the community (Fig. 2(g,h)).

Kronecker graphs are analytically tractable as we can prove [5] that they follow a range of properties also found in real networks: heavy tails for in- and out-degrees, clustering coefficient, power-law spectrum. They also densify and have shrinking diameter.

One can interpret the diagonal values of the initiator matrix as the proportion of edges inside each of the groups, and the off-diagonal values give the fraction of edges connecting the groups (Fig. 2(h)). So, for networks with hierarchical community structure the diagonal elements have higher values than off the diagonal elements, *i.e.*, communities have more edges internally than edges pointing between the communities. However, as we estimate [6] the Kronecker initiator matrix from real networks we typically obtain matrices as in Fig. 2(g), where the top left element is the largest and then the values on the diagonal decay faster than off the diagonal.

Again this suggests a *core-periphery* (Fig. 2(d)) network structure where the network is composed denser and denser layers, like an onion. In case of Kronecker graphs the core is modeled by the top-left entry and the periphery by the bottom-right (Fig. 2(g)). Most edges are inside the core (large top-left), and very few between the nodes of periphery (small bottom-right), while there are many more edges between the core and the periphery than inside the periphery (relatively large off-diagonal) (Fig. 2(g)). And in spirit of Kronecker graphs the structure repeats recursively — the core again has the dense core and the periphery, and so on. And similarly the periphery itself has the core and the periphery. This suggest an “onion” like network structure as illustrated in Fig 2(h), where the network is composed of denser and denser layers as one moves towards the center of the network.

Interestingly, this nicely connects with the observations from the shape of NCP plot. Either by using the graph partitioning approach to quantify the community structure of large networks or by using a Kronecker graphs model and fitting it to real data, we arrive to the same conclusion. In contrast to small networks that often ex-

hibit unambiguous community structure, large networks organize themselves into a nested core-periphery structure with small communities on the periphery and denser and denser layers of the core.

4. DISCUSSION AND CONCLUSION

As communities tend to exist only at small size scales of up to about 100 nodes, while at larger size scales network communities become less community-like, our observations also demonstrate that attempts at community identification in large networks will be particularly difficult in the network core and when one focuses exclusively on the connectivity structure of the network.

One of the implications of this is that networks do not break nicely into hierarchically organized sets of communities that lend themselves to graph partitioning and community detection algorithms. On contrary, this suggests that large networks can be decomposed into a densely linked core with many small periphery pieces hanging off the core. It is an intriguing question whether there exist graph-based community detection methods that could effectively recover such groups of nodes.

Acknowledgements

Based on joint work with D. Chakrabarti, A. Dasgupta, C. Faloutsos, Z. Ghahramani, J. Kleinberg, K. Lang and M. Mahoney.

5. REFERENCES

- [1] R. Andersen, F. Chung, and K. Lang. Local graph partitioning using PageRank vectors. In *FOCS*, 2006.
- [2] A. Clauset, C. Moore, and M. Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98–101, 2008.
- [3] S. Fortunato. Community detection in graphs. *Arxiv*, 2009.
- [4] M. Gaertler. Clustering. In U. Brandes and T. Erlebach, editors, *Network Analysis: Methodological Foundations*, pages 178–215. Springer, 2005.
- [5] J. Leskovec, D. Chakrabarti, J. Kleinberg, and C. Faloutsos. Realistic, mathematically tractable graph generation and evolution, using Kronecker multiplication. In *ECML*, 2005.
- [6] J. Leskovec and C. Faloutsos. Scalable modeling of real graphs using kronecker multiplication. In *ICML*, 2007.
- [7] J. Leskovec, K. Lang, A. Dasgupta, and M. Mahoney. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. arXiv:0810.1355, October 2008.
- [8] J. Leskovec, K. Lang, A. Dasgupta, and M. Mahoney. Statistical properties of community structure in large social and information networks. In *WWW*, 2008.
- [9] S. Schaeffer. Graph clustering. *Comp. Sci. Review*, 2007.