18-859: Special Topics in Communications: Information-Flows in Networks **Fall 2014**

Lecture 8: Network Science I

Lecturer: Pulkit Grover and Osman Yagan *Scribes: Anit Sahu and Yaoqing Yang*

Note: *LaTeX template courtesy of UC Berkeley EECS dept and CMU ML dept.*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

8.1 Motivation

8.1.1 Why do we study networks?

Network science might be one of the most interdisciplinary sciences. It has huge impacts on many areas such as computer science, power systems, biology and social science. In what follows we give some examples about network science.

- Social networks (Epidemics, viral marketing, cultural fads...);
- Internet (Robustness against failures and attacks, routing policies and protocols, optimizing the future growth of Internet...);
- The airline networks (Optimization of capacity and resource consumption under extreme reliability requirements...);
- Power grid (Mitigating failures, Optimizing generation/transmission/distribution);
- Protein interaction network (Drug design, Gene theory, biomarkers of a disease);
- Software Systems (Function calls, detecting the sources of bugs).

It is necessary that a cross-cutting science of networks be establishes during the 21st century.

8.1.2 What do we study in network science?

A network can often be represented by a graph. Therefore, we often use tools from graph theory to represent and analyze networks. A graph can be denoted by $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of nodes and $\mathcal{E}$ is the set of edges. Suppose $|\mathcal{V}| = n$, then the graph $\mathcal{G}$ can also be represented by a zero-one adjacency matrix $A = (a_{ij})$, where $a_{ij} = 1$ when nodes v_i and v_j are connected. In the following we review some basic coefficients and properties of a graph.

8.1.2.1 Degree Distribution

The degree $\deg(v)$ of a node $v \in V$ is the number of edges that connects with v. The degree distribution usually refers to the empirical degree distribution of the node degree. That is, if the total number of nodes is n and the number of nodes with degree k is n_k, the distribution function $P(k)$ is written as n_k/n.

The degree distribution can also mean the probabilistic distribution of the node degree. For example, in a random network with n nodes and each two nodes being connected with some probability p, the degree distribution $P(k)$ can be written as

$$P(k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}. \tag{8.1}$$

Different networks tend to have different degree distributions. For example, the Internet has a degree distribution that follows a power law $P(k) = k^{-\gamma}$ but some other networks, like data center networks, tends to have highly unbalanced degree distribution. Therefore, network degree distribution is a useful tool to select from different network models.

8.1.2.2 Clustering Coefficient

The clustering coefficient C is defined as

$$C = \frac{3 * \text{No. of triangles}}{\text{No. of triplets}} = \frac{|\{\{i,j,k\}|\{(v_i,v_j),(v_j,v_k),(v_i,v_k)\} \subset E\}|}{|\{\{i,j,k\}|\{(v_i,v_j),(v_i,v_k)\} \subset E\}|}. \tag{8.2}$$

When a graph is complete, the clustering coefficient is 1. This parameter characterizes the tendency of the nodes in the graph being clustered together. An intuitive explanation of clustering coefficient in a social network is the tendency that one's friends are friends of each other.

8.1.2.3 Connectivity and Components

A network is connected if any two nodes are connected via a finite chain of other nodes. If we want to further describe whether the network is highly connected or not, we have to introduce the k-connectivity. If a network is k-connected, then there are at least k mutually disjoint paths between each pair of nodes. An equivalent definition is that, a network is k-connected if the network remains connected when we arbitrarily removes k-1 nodes.

When a network is not connected, we use the size of components to study the extend of being unconnected. A component means a connected subgraph. The component with the maximum number of nodes is called the largest component, the size of which is denoted by $C_{\max}$. The existence of large components are studied extensively in random graph theory [2]. Studying large components is important for studying epidemics in networks.

8.1.2.4 Diameter

The distance of two nodes v_i and v_j is defined as the minimum number of hops that v_i are connected to v_j. The diameter of the network is the largest distance among all node pairs (v_i, v_j). A famous example is that any two people in the world might be connected through a chain of approximately 6 people in the network of acquaintance [1]. Diameter is crucial in data transmission, since it characterize the worst case of transmitting data from one end of the network to another end.

8.1.2.5 Centrality

The term centrality in network science has variant definitions. Basically, centrality of a node characterizes the importance of this node in the whole network. The definition given in the class is related to the information flow: the centrality of a node is the fraction of shortest paths passing through a node. Under this definition, a node with high centrality tends to be the bottleneck of flow transportation.

8.1.2.6 Assortative Mixing

By assortative mixing, we denote the extend to which nodes with similar attributes are likely to link to each other. The social network is certainly a good example of high assortative mixing networks, since a person is more or less interested in having friends with the ones that have similar age and hobbies as himself.

The above parameters and properties are more or less related to describing the topology of a network. Apart from topology, we can also study the various processes and behaviors of a network. A core issue is network flow. In the following, some examples of network flow are given

- spread and epidemic of diseases;
- routing data;
- propagation of tweets;
- materials transport/flow;
- gossip spreading and marketing;
- cascade of failures.

8.1.3 How do we study networks?

Graph theory provides the general toolbox for studying networks. In 1735, Euler initiated the study of graph theory by solving the problem of Seven Bridges of Königsberg. The problem is to find out a route that passes through each one of the seven bridges in Königsberg once and only once. Euler abstracted each bridge as an edge of a graph and observed the existence of more than two odd-degree nodes, which proved the impossibility of constructing a required route.

The foundation of modern graph theory is given by Paul Erdös and Alfréd Rényi in a series of papers on random graphs. Two basic models on random graphs are $\mathcal{G}(n;p)$ and $\mathcal{G}(n;M)$. The random graph $\mathcal{G}(n;p)$ has n nodes and each pair of nodes are connected with probability p. All connections are independent from each other. The random graph $\mathcal{G}(n;M)$ has a uniform distribution over all graphs with n nodes and M edges. Suppose $G(n;p)$ and $G(n;M)$ are respectively instances of $\mathcal{G}(n;p)$ and $\mathcal{G}(n;M)$, then $G(n;p)$ and $G(n;M)$ might have very similar properties if $p\binom{n}{2} = M$.

In the initial study of random graphs, Erdös and Rényi focus on the existence of certain components $\mathscr{A}$ in a random graph. $\mathscr{A}$ might be as simple as a triangle or as complicated as a giant component with some topology constraints. Particularly, Erdös and Rényi are interested in finding the limit

$$\lim_{n\to\infty} \Pr(G(n;p_n) \text{ has } \mathscr{A}), \tag{8.3}$$

where p_n is a function that maps n to $[0,1]$. They found that in random graphs, a phase transition phenomenon exists. That is, there exists a threshold $T(\mathscr{A})$ such that when the parameter of p_n crosses $T(\mathscr{A})$, the limit value in (8.3) changes from 0 to 1 drastically. In what follows, we show to classic results on phase transition. The acronym w.h.p. means *with high probability.*

Theorem 1 Suppose $p_n = \frac{c}{n}$ and c is a constant, then

- If $c < 1$, then w.h.p. $|C_{\max}(p_n)| = O(\log n)$;
- If $c > 1$, then w.h.p. $|C_{\max}(p_n)| = \beta(n + o(n))$ where $\beta + \exp(-\beta c) = 1$.

Theorem 2 Suppose $p_n = \frac{c \log n}{n}$ and c is a constant, then

- If $c < 1$, then w.h.p. the network is not connected;
- If $c > 1$, then w.h.p. the network is connected.

8.2 Small World Network

Although the $G(n;p)$ and $G(n;M)$ models are studied extensively, they might not be able to capture real world networks. For example, a social network often has small diameter (e.g. the six degrees of separation) and high clustering coefficient. However, although the $G(n;p)$ and $G(n;M)$ models also have small diameter which is in the order $O(\log N)$, these two models generally have low clustering coefficients. One way to solve this discrepancy is to combine these network models with another simple network model, the regular lattice.

First, we construct a ring with n nodes. Then we connect each node v to k nearest nodes on both sides of v. This scheme can provide a circular lattice with high clustering coefficient. After that, we rewire each node's leftmost connection with probability p and connect this wire to other nodes uniformly. After k iterations [3], we create a small world network if p is properly chosen. In fact, if p is too high, the clustering coefficient drops drastically, and if p is too low, the diameter of the graph is still high (because a circular lattice has a high diameter). Therefore, a small world network can only be created when p is moderate.

8.2.1 An example of Epilepsy

In the class, we saw an application of small-world network in the study of epilepsy. Epilepsy was believed to be caused by hypersynchronous neuronal activity. However, in [4], the authors address epilepsy from a pure network epidemic perspective.

The motivation to restudy epilepsy from a pure network-science perspective are the discrepancies of evidences from hippocampal slices from CA3 and CA1 regions. The CA3 and CA1 are respectively the places that burst-type epilepsy and the seizure-type epilepsy originate. However, evidence from CA1 region shows that neuronal activities during seizures are not synchronous, which is inconsistent with former belief of epilepsy. Moreover, the most significant difference between CA3 and CA1 regions is the fraction of long-distance synaptic connections. Therefore, the authors conjecture that the two different types of epilepsies are caused by the same principle, but appear different because of the wiring patterns of different brain regions.

Two small world networks based on random rewiring are then created to simulate two different kind of epilepsies. The difference in the small world networks is only the rewiring probability. The first network has a lower rewiring probability and a seizure-type epidemic is observed, while the second network has a higher rewiring probability and a burst-type epidemic is observed. The explanation is that in the network with high rewiring probability, more long-distance connections result in more new firing events. The firing events happen too quickly and the whole network fires simultaneously, which leads to a burst. After this burst, all neurons enter a refractory period and recover simultaneously. However, when the rewiring probability is low, the firing events slowly spread to the whole network, but some part of the network recover and become the new agents that can be fired. This results in a long-lasting seizure-type epidemic, which is mild when comparing with burst-type epidemic.

References

[1] Stanley Milgram. "The small world problem." Psychology today 2, no. 1 (1967): 60-67.

[2] Paul Erdös and Alfréd Rényi. "On the evolution of random graphs." Publ. Math. Inst. Hungar. Acad. Sci 5 (1960): 17-61.

[3] Duncan J. Watts and Steven H. Strogatz. "Collective dynamics of small-worldnetworks." nature 393, no. 6684 (1998): 440-442.

[4] Theoden I. Netoff, Robert Clewley, Scott Arno, Tara Keck and John A. White. "Epilepsy in small-world networks." The Journal of Neuroscience 24, no. 37 (2004): 8075-8083.