Social and Technological Networks Edinburgh, 2018

Notes 5. Clustering and community detection.

Rik Sarkar *Notes*

Network Flows and application to community detection

One of the techniques for community detection that we have not discussed in class is the use of network flows. This is a major and important area of graph theory.

The ideas is as follows. Suppose, for every edge e in the network, we are given a capacity, given by the function $c(e)$. The capacity may represent the width of pipe, and therefore how much water can flow through it per unit time. Or, it may represent the width of road, or speed of a computer network connection etc. Such a network is called a capacitated network.

Now suppose the network has has two nodes – source s and target t. The network flow problem is: *What is the maximum possible flow between s and t that the network can support?*

A related notion is an $s - t$ cut. *An $s - t$ cut in the graph is a set of edges whose removal disconnects s from t.* The *minimum cut is the set of edges with smallest total capacity that is a cut.*

Theorem 0.1 (Max flow – min cut). *The maximum $s - t$ flow is given by the capacity of the minimum $s - t$ cut.*

You can find this theorem on the wikipedia page: `https://en.wikipedia.org/wiki/Max-flow_min-cut_theorem`. The book CLRS: Introduction to Algorithms has more details on network flows if you are interested. Algorithms finding the max flow tend to be $O(n^3), O(nm^2)$ and similar complexities.

The idea behind the theorem is quite intuitive. Imagine that there are tight communities around s and t, while the connection between them consists of one or few weak (low capacity) links. Then, however dense other sets of edges may be, this bottleneck of weak links will restrict the flow down to their capacity. Therefore, this is a simple method to determine two communities where one contains s, the other contains t, and the cut between them.

Optional reading: multiple communities using flows. [Source: Kempe 2018] In the more general case when we are looking for multiple communities, we can run network flow for every pair of vertices. The complexity of this can be reduced using techniques like *Gomory-Hu trees.* These are trees of cuts in the graph representing relation between all cuts in the graph.

This tree has $n - 1$ vertices, each corresponding to a cut, and can be used to find sets of multiple cuts/communities in the graph.

The problem of finding densest subgraph (the one with larges edge-density) can be solved using another version of max flow algorithm by Goldberg and Tarjan.

Community structure in random graphs.

We have discussed communities in various graphs. But what about random graphs, our base case – the prototype? Intuitively, random graphs should not have communities, since edge to a node outside the community is as likely as an edge inside the community.

Let us try to prove this a bit more specifically. Suppose we are interested in communities of size about αn for some fraction α. Also, let us go a little bit extreme, and say that a graph has community structure if there are two sets of αn nodes with no links between them. This is not that unrealistic. In a graph with several communities (imagine α is something small, like 0.01), it is quite possible that there communities with no direct links.

Suppose $G = (V, E)$ is an ER graph with $p = c/n$, and $S, T \subset V$ are two given sets of αn vertices each.

Q 1. *First of all, show that the probability that there are no edges between S and T is $\leq e^{-c\alpha^2 n}$.*

Q 2. *Now, check that the number of possible choices of S and T is at most $\binom{n}{\alpha n}^2$.*

Q 3. *Now show that, if $c > 2\ln(e/\alpha)/\alpha$, the probability that there are no two communities S and T with an edge between them tends to zero as n grows.*

For the solution to this, see Dan Spielman's notes, Sec 3.6.

More generally, we can expect that the probability of there being few edges relative to the density of edges in the graph between different subsets of the graph are also small. But this will require a more complex analysis.

Other notes.

Q 4. *How fast can the edge density of a subset $S \subset V$ grow? Suppose we use notion $n = |V|$ and $x = |S|$.*

Q 5. *Give example of two graphs A, B, such that A contains a smaller fraction of possible edges than B, but has greater density.*