Community Detection

Joseph McGuire

11/15/2021

Setup

Networks

Network's are common ways of visualizing and representing large data. It's common for networks to become quite large very quickly:

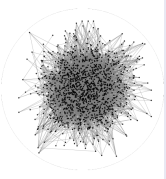

Figure: (Abbe, 2017) [1]

This can represent anything from social interactions, to stock trades, to biological processes.

Motivating Example

Over a 2 year period sociologists closely followed a subject, Zachary, to their Karate club of 34 people. Friendships were recorded amongst the members were marked as edges, with individuals being vertices, resulting in the following network:

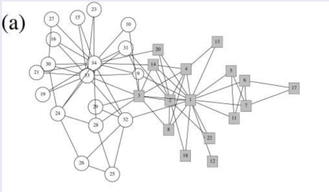

Figure: (Newman,2002) [6]

One might desire to identify factions/friend circles and suggest the growth of these factions by recommending two people meet. Think Facebook friend recommendations.

Community Detection: Community

Community Detection

The goal of community detection is to partition a network in such a way as to find the subgraphs that are internally highly connected and externally sparsely connected.

Definition (Community)

Let $G = (V, E)$ be a simple graph, with degree n. A community of $H = (V_H, E_H) \subset G$ is a subgraph of G, such that for any $h \in V_H$, let h_i be the number of internal connections of h in H (that is edges of h in E_H) and h_e be the number of external connections of h (that is, edges of h in $E - E_H$), then

$$h_e << h_i.$$

This can be generalized to weighted graphs, by considering sums of weights instead of number of connections.

Community Detection: Goal

Goal

Without a priori knowledge of these communities, this is a difficult problem to solve. The goal is to recover structures like this:

Figure: (Abbe,2017) [1]

Community Detection: Types of Algorithms

There are many different approaches on how to do this, one way to classify these algorithms is by their domain of consideration:

- Node-Centric, seek to solve the problem by rules imposed on individual nodes;
- Network-Centric, these tend to focus on the overall network topology;
- Hierarchy-Centric, these seek to reconstruct the network in such a way that reveals the hierarchical structure.

Community Detection: Node-Centric

k-Distance Communities

One classification technique is by determining communities via paths; for example, we might consider two nodes in the same community if they can be connected by a $k-$length path.

Formally , we can use this intuition to define a k-community V_s as the set such that:

$$d(v_i, v_j) \leq k \qquad \forall v_i, v_j \in V_s,$$

where $d(x, y)$ is some distance function and k is some upper-bound of this function.

Community Detection: Network-Centric

Latent Space Models

Latent space modeling takes the graph partition problem of graphs and maps it to Euclidean space.

That is take your nodes V from a weighted graph G and map them into a Euclidean space using a map determined by the shortest walk between two points. One such embedding is seen in the next slide:

Community Detection: Network-Centric

Latent Space Models

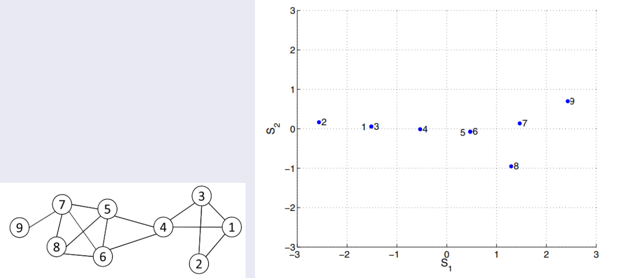

Figure: (Tang and Huan, 2010)[3]

From here, the method of $k-$means clustering can be employed to determine community structures in the new Euclidean space. We then map this back into our graph to determine our communities.

Community Detection: Hierarchy-Centric

Modularity Optimization

Hierarchical models focus on producing a hierarchy of nodes typically through an optimization of some objective function.

That is, some function of the graph and its communities that identifies community structure in some way.

Community Detection: Modularity

Definition (Modularity)

For a weighted graph $G = (V, E)$ with an adjacency matrix A with N communities, we define the modularity of G:

$$Mod = \sum_{i=1}^{N}(e_{ii} - a_i^2) \qquad Q \in [-\frac{1}{2}, 1]$$

Where

- e_{ii} is the fraction of edges that end in the same community c_i; $e_{ii} = \sum_{v,w \in c_i} \frac{A_{v,w}}{2m}$
- $2m = \sum_{i,j} A_{i,j}$;
- a_i is the fraction of edges that are attached to a vertex in c_i.

This is measure of the clustering of the communities of G, a high *Mod* value, means we have a good community structure; meaning the partitions are highly internally connected, and sparsely externally connected.

Community Detection: Modularity

Examples of Modularity

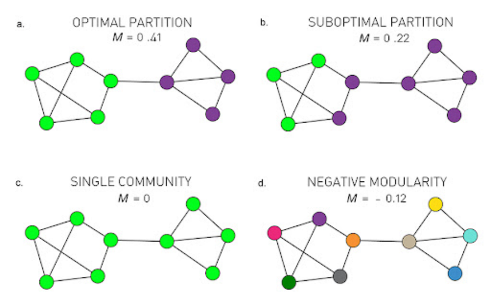

Figure: (Barabasi, 2013) [5], $M = Mod$

Community Detection: Modularity Maximization

Modularity Maximization Algorithm

From 4, this is a modified Girvan-Newman model. For a non-directed weighted graph $G = (V, E)$ with $|V| = N$.

- Phase 1:
  1. (Initialization)
     Each node is assigned to one community; that is, every node v_i is assigned to community c_i. Pick some starting node v_1.
  2. (Modularity Testing)
     For each neighbor v_j, calculate ΔMod for each v_i; that is, consider moving v_i from c_i and placing it in c_j, where c_j is the community of v_j.
  3. (Modularity Optimization)
     Place v_i in the community c_j such that ΔMod is maximized. If necessary, employ a tie-breaking rule.
  4. (Loop)
     Repeat for node v_{i+1}.
  5. (Terminate) Continue iterating over the cells, with possible repeats, until $\Delta Mod = 0$ for all nodes v_i.

Community Detection: Modularity Maximization

Modularity Maximization Algorithm

- Phase 2:
  1. Initialize a non-directed weighted graph $G' = (V', E')$ where each community C_j is now a node in V' and the edges and their weights are determined below;
  2. The sum of the inter-connected weights inside the community is weight of the self-loop;
  3. The weights of the external-edges, between the nodes, is determined by the sum of the connections between the two communities;
  4. Save G', and repeat phase 1 with G'.

References

1. Abbe, Emmanuel. "Community detection and stochastic block models: recent developments." The Journal of Machine Learning Research 18.1 (2017): 6446-6531.
2. Newman, Mark EJ. "Analysis of weighted networks." Physical review E 70.5 (2004): 056131.
3. Tang, Lei, and Huan Liu. "Community detection and mining in social media." Synthesis lectures on data mining and knowledge discovery 2.1 (2010): 1-137.
4. Blondel, Vincent D., et al. "Fast unfolding of communities in large networks." Journal of statistical mechanics: theory and experiment 2008.10 (2008): P10008.
5. Barabási, Albert-László. "Network science." Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 371.1987 (2013): 20120375.
6. Girvan, Michelle, and Mark EJ Newman. "Community structure in social and biological networks." Proceedings of the national academy of sciences 99.12 (2002): 7821-7826.