Networks and Community Detection

Joshua Mankelow

February 16, 2022

Abstract

Networks are a great model for analysing and understanding real world complex systems. A very important feature of these network models is *community detection*. Community detection allows us to identify clusters in the network that are well connected amongst eachother. If a network is being used to model a real world system then finding this structure has many implications about the behaviour of the system. This essay will discuss multiple methods for community detection in networks and their applications to the analysis and understanding of real world complex systems.

Contents

Todo list

KEY: NOT STARTED

KEY: IN PROGRESS

KEY: DRAFT FINISHED

KEY: IN REVIEW

KEY: COMPLETED

KEY: QUESTION

========

1 Introduction to Networks

SEC: Introduce Network Theory

Networks in the technical sense are analogous to networks in the nontechnical sense - a collection of objects paired with a number of connections that can link any two objects. *Networks: An Introduction* by *Mark Newman* lists *Technological Networks*, *Social Networks*, *Information Networks* and *Biological Networks* as different systems that are modelled by the technical interpretation of a network.[New10, Contents]. A brief example of a network would be something like the following: Imagine you and a number of people you speak to regularly are represented as dots (nodes or vertices) on a piece of paper. Then if any two people are friends, the dots representing those people are connected by a line (edge). If you then repeat this process by asking your acquaintances to list all their friends and so on, you will end up with a simple model of a *social network*.

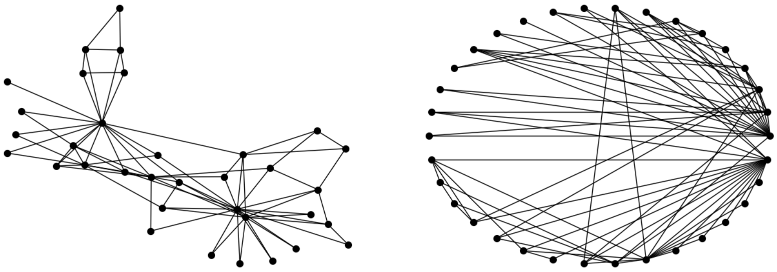
Now that we have this model it's easy to be curious about any structure that emerges that we can detect and abuse to develop an understanding of the real world system that we are representing. The structure that this essay will explore is that of *communities*. Vaguely speaking, communities are subsets of a network that are *densely connected* amongst themselves. I.e. there is some notion of any node within a community being more closely connected to other nodes in the community than nodes outside the community in the average case. Before we dive into the details of communities and detecting them, I wish to provide some motivation by way of example of the kinds of situations that networks can arise and why they are the natural model for the related systems.

1.1 Social Networks

SEC: Models of Networks

To better illustrate the simple notion of a social network mentioned above, I will introduce the canonical community detection example of *Zachary's Karate Club*. Zachary's Karate Club is a dataset where "The data was collected from the members of a university karate club by Wayne Zachary in 1977. Each node represents a member of the club, and each edge represents a tie between two members of the club."[kon17, Metadata]. In Figure 1, there are two different renderings of the Zachary Karate Club. Figure 1a shows the network rendered using a "spring" layout (which is a type of force directed graph drawing[Kob12])and figure 1b shows the network rendered using a "circle" layout. These different layouts show us different parts of the underlying structure of the network. For example, in Figure 1a, it's clear which nodes in the network have the highest degree and which are of lower degree. It also allows you to see some of the community structure in the network. Meanwhile, in Figure 1b, it's much easier to see the which nodes edges in the network would need to be removed to disconnect the network in a minimal way. The reason this dataset is the canonical example of community detection is that the question that comes with it is the following: Suppose two members of the club have a disagreement which causes the club to split in two. How does the club split? In Zachary's original paper on the topic *An Information Flow Model for Conflict in Small Groups*[Zac77] he uses community detection techniques to predict how the network will split after the disagreement. Out of 34 people, Zachary correctly predicts how 33 of them will choose a side after the disagreement.

There, of course, exist different ways to represent social networks. The way



(a) Spring Layout

(b) Circle Layout

Figure 1: Two renderings of the Zachary Karate Club network using data from KONECT.cc[Kun13] and a Python library NetworkX[HSS08]

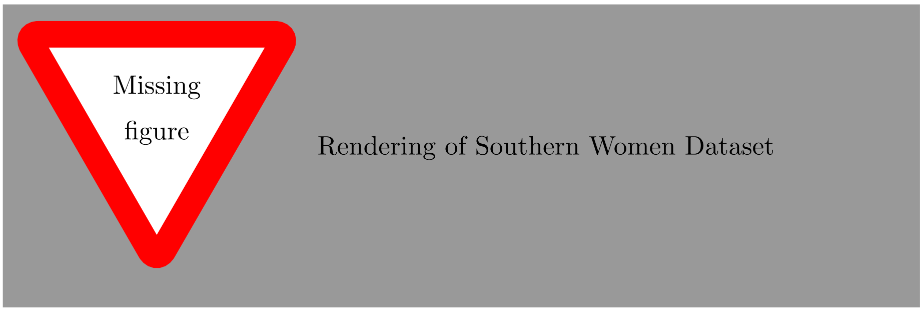


Figure 2: A rendering of the Southern Women Dataset

in which you choose to represent them depends on the question you're trying to answer. For example, one might imagine having two types of nodes in a network. One type of node will represent a person and another type of node will represent an event. An edge is drawn between a person and an event if a person attended a given event and person A is considered connected to person B if they both attended the same event. One such example of this is the *Southern Women Dataset.*[DGG41] This dataset is another example of a community detection problem because after analysis of the data, it was found that women in the group were split into two discrete subgroups.

1.2 Technological Networks

As a result of our intensely and digitally connected world, technological networks are of significant interest to researchers. The easiest example to consider is the Internet. The internet consists of many computers all connected by copper or fibre cables which signals are sent through to transmit data. As one might imagine, in the model, the computers are nodes and the cables are the edges. The internet needs to be robust against software and hardware failures and this is where the idea of commnity detection can help us. Saying that we want the internet to be robust is the same as saying that we want every node in the

Figure 3: A rendering of the Internet

network to be strongly connected to every other node i.e. the number of possible routes between any two nodes is large. This means that, in the philosophy of community detection, we want the internet to act as one large community rather than multiple smaller communities that are loosley connected. An alternative way of looking at this is that once we've managed to identify the communities, we can then figure out which edges and noes are the critical ones that allow passage from one community to another. This allows us to reinforce those edges and nodes to reduce the potential for failure.

Yet another example of a technological network would be the UK Power Grid. Network theory is a useful model here as with the UK Power Grid we're trying to solve exactly the same problem as with the internet — we want the system to be robust against hardware or software failures.

1.3 Information Networks

The easiest example of an information network is that which is generated by looking back through the citations on a paper recursively. If Paper A references Paper B, then we will draw a directed edge connecting Paper A to Paper B. This will generate a network that shows which papers are referenced by which other papers and how information is reused. Applying community detection to such a network would show us the different academic working groups and perhaps even different fields or subfields of a subject.

Another example of an information network is the World Wide Web which differs from the internet in that it refers to the webpages hosted on the internet rather than the servers and cables themselves. Mapping the world wide web as a network shows us communities of websites that regularly reference each other.

2 Properties of Networks

SEC: Definition of a Network

Community detection relies on us knowing lots about the underlying structure of a network and to do that we have to understand its properties. This chapter will establish a more formal understanding of networks and will highlight some key properties and methods that we will use to exctract value about community structure later.

Definition 1. *(Undirected network) Let V be a set of vertices (nodes) and let E be a set of pairs of vertices such that if $e = (x, y) \in E$ then $x, y \in V$. An undirected network is the pair $(V, E) = N$. An edge $e = (x, y) \in E$ is said to join x and y and y to x.*

The undirected network is the simplest type of network and on its own has interesting enough properties. However, for the sake of example and application, we will also introduce some other types of network that allow for more detailed models.

SEC: Different Types of Network

Definition 2. *(Directed network) Let V be a set of vertices (nodes) and let E be a set of pairs of vertices such that if $e = (x, y) \in E$ then $x, y \in V$. A directed network is the pair $(V, E) = N$. An edge $e = (x, y) \in E$ is said to join x to y. I.e. if x is joined to y then y is not necessarily joined to x.*

The intuition for directed graphs, is that edges may only be travelled along in one way. This comes in handy for modelling more intricate systems. The final network type of interest is that of the weighted network.

Definition 3. *(Weighted network) Let V be a set of vertices (nodes) and let E be a set of triples of the form $V^2 \times \mathbb{R}$ such that if $e = (x, y, w) \in E$ then $x, y \in V$. The value w is said to be the weight of the edge.*

SEC: Interesting Properties of Networks

The weighted network allows us to introduce some notion of how hard it is to move along a certain edge. This is useful when modeling things like traffic flow. [citation needed]

2.1 Adjacency Matrices

The objects defined above are meaningless without a concrete way of mathematically representing them. To that end, we have to come up with a way of describing a network mathematically. This leads us to the definition of the adjacency matrix:

Definition 4. *(Adjacency matrix) Let $N = (V, E)$ be a network and label every vertex $v \in V$ with a number from 1 to $n = |V|$. The adjacency matrix of a network is the matrix of elements $(A)_{ij}$ such that $a_{ij} = 1$ if $(i, j) \in E$ and $a_{ij} = 0$ if $(i, j) \notin E$. In other words, if nodes i and j are connected by an edge in the network, then the corresponding element in the matrix is 1. Otherwise, it is 0.*

The adjacency matrix gives us our first way of representing a network. This will form the basis for most of the analytical work we do going forwards. It's worth noting that there are also different types of adjacency matrix corresponding to the different types of network. For example, in the case of a directed network we will have a non-symmetric matrix where $a_{ij} = 1$ if $(i, j) \in E$ but this does not necessarily mean that $a_{ji} = 1$. We also get something similar for weighted networks where we set $a_{ij} = w$ the weight of the edge connecting i and j in N.

2.2 Bipartite Graphs

Not really sure if it's worth talking about these.

2.3 Paths

When we're analysing a network, we're very often interested in which vertices are reachable from any given vertex. As such, we become interested in the idea of a path. A path in a network is defined in the following way

Definition 5. *(Path) Let $N = (V, E)$ be a network. A path is a sequence of distinct vertices $v_1, \ldots, v_n \in E$ such that $(e_i, e_{i+1}) \in E$ for all $i = 1, \ldots, n-1$. In other words, a path is a sequence of distinct vertices such that every consecutive pair of vertices is connected by an edge in E.*

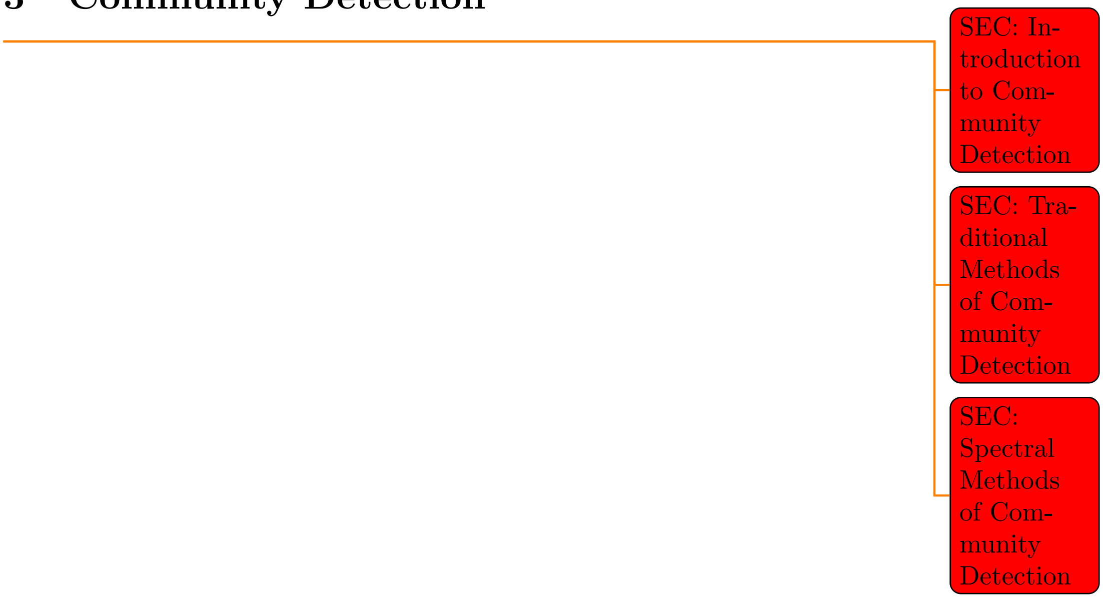
Paths are an important concept in community detection as they allow us to phrase questions in rigorous terms as opposed to loose concepts of connectedness.

2.4 Components

2.5 Cut Sets

2.6 The Graph Laplacian

3 Community Detection


SEC: Introduction to Community Detection
SEC: Traditional Methods of Community Detection
SEC: Spectral Methods of Community Detection

4 Applications of Community Detection

SEC: Applications of Community Detection

SEC: Figure out an interesting thing to write some of my own code for

References

[DGG41] Allison Davis, Burleigh B. Gardner, and Mary R. Gardner. *Deep South; a Social Anthropological Study of Caste and Class.* The Univ. of Chicago Press, 1941.

[HSS08] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using networkx. In Gaël Varoquaux, Travis Vaught, and Jarrod Millman, editors, *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA, 2008.

[Kob12] Stephen G. Kobourov. Spring embedders and force directed graphs. 2012.

[kon17] Zachary karate club network dataset – KONECT, October 2017.

[Kun13] Jérôme Kunegis. KONECT – The Koblenz Network Collection. In *Proc. Int. Conf. on World Wide Web Companion*, pages 1343–1350, 2013.

[New10] Mark Newman. *Networks: An Introduction.* 2010.

[Zac77] Wane W. Zachary. An information flow model for conflict in small groups. 1977.