Ain't Nothin But a G-Trie

A Thesis
Presented to
The Division of Mathematics and Natural Sciences
Reed College

In Partial Fulfillment
of the Requirements for the Degree
Bachelor of Arts

Arthur James Lawson III

May 2022

Approved for the Division
(Computer Science)

James D. Fix

List of Abbreviations

You can always change the way your abbreviations are formatted. Play around with it yourself, use tables, or come to CUS if you'd like to change the way it looks. You can also completely remove this chapter if you have no need for a list of abbreviations. Here is an example of what this could look like:

AI	Artificial Intelligencet
CPU	Central Processing Unit
GFD	Graphlet Frequency Distribution

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Millions of people use social media every day. The amount of data available from this casual use is mind-numbingly large and can be used to serve important purposes such as training artificial intelligence (AI), making more personalized advertisements, and various forms of analysis of human social behavior. As social media becomes more prevalent, concerns around privacy are growing. In an ideal world, we can use this data in a way that can help us learn, grow, and improve. Before we do that, we take a look at how graphlet census helps identify the structure of a network.

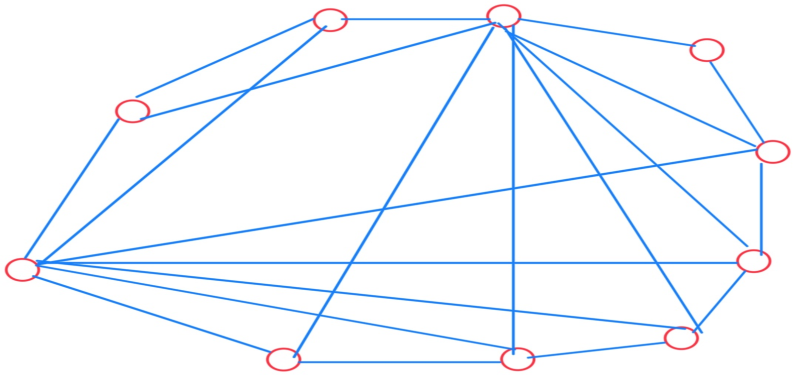
Have you ever wondered how anonymous a network truly is? Many studies we see online talk about data being safe and ethical because it has been anonymized, but what does anonymous mean in this context and how do we measure it? Is data safe because it replaces names with serial numbers? Is it possible to work backwards from an ”anonymous” network and figure out what the original data set was? These are the questions that have guided my research over the past (–n–) months.

1.1 Social Networks

Let’s take a look at Facebook. The basic structure is people connecting with other people. The concept works because a user knows that when they create a profile, they can easily find and connect with their friends, family, and colleagues. A lot of information can be taken from even small friend networks.

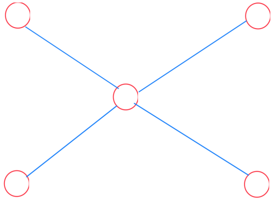
For example, a friend network of 3 people could tell you a lot. In the case where there aren’t any connections, it is pretty safe to assume that this social network won’t be very successful. Each person that posts can only see their own posts and this site for sharing with others quickly becomes a diary with additional complications (such as not working offline). If 2/3 of these people are friends, both of those two now have much more reason to use this application. Instead of talking to themselves, they are interacting with another person (being social!). Now suppose that the three people form a triangle. This means that they are all friends and become a lot more likely to post and interact because they know that they have two different people that could interact with their posts. They no longer log on to see only what person A had for breakfast, but they walk into a virtual world where person B has commented an unbelievable anecdote about eating that exact same cereal across from The Weeknd.

Figure 1.1: Can you pick the professor?



As networks increase in size, the possible shapes, and their implications, grow in complexity and potential value. For example, see Figure 1.1. In a social network of 10 Reed CS community members, there are a lot of questions to be asked. If I told you the student to professor ratio was 8:2, could you pick out who the two professors are without peeking at their long list of degrees? How about you look at each node's degree (–can turn this into a much more clever joke later –). After a look at the network and one would notice that two of the nodes have a lot more connections than the rest. It's a safe assumption that the two professors are the people with the high number of connections because students are very interested in their wisdom, humor, and pictures of Eitan flying his plane!

Figure 1.2: Spot the tutor?



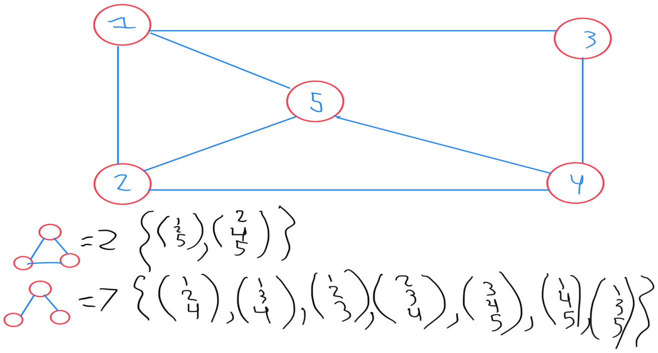
Similarly, what if I asked you to identify the 121 tutor in a group of 5 Reed students (Figure 1.2). Well, if you notice that there is only one person connected to every other, it would be a safe bet to say the person in the middle is the tutor. (With image, include definition of a star. OR can do so in later reference when we are defining shapes)

1.2 Subgraph Counting

Analysis of connections within a large network can be simplified with a bottom-up approach. Tens of millions of people in the US alone use Facebook on a daily basis.

How can you analyze millions of people, and the billions of connections between them? You break the larger problem into a smaller problem –insert- parallel joke– . In a network of 5 people, you can focus on a smaller network of 3 people. The network of 3 people is a *subgraph* of the original network. We will provide a more formal definition for subgraph in the next chapter. The focus of this thesis is tackling the subgraph counting problem which can be defined as: given a specific collection of subgraphs and a network, how many times does each subgraph occur in the larger network?

Figure 1.3: Subgraph Counting



In Figure 1.3 we see an example of subgraph counting within a larger network. A single vertex can participate in different subgraphs as long as at least one of the others is different (such as node 1 participating in multiple subgraphs).

1.3 Motifs

In any network, there are certain patterns that are bound to become recurring characters. *Network motifs* are reoccurring subgraphs with statistical significance. So far we've discussed triangles and stars (a graph term to represent shapes such as that of Figure 1.2 where one person is connected to everyone else). Let's take a look at some more specific shapes that can occur in a network.

In Figure 1.4, there are 10 shapes that represent the possible combinations of 1, 2, 3, and 4 objects within a network. The primary work of graph tries (G-trie) is to store these shapes in a tree-like structure.

See Fig 1.5 for an example of a *prefix trie*. At each level, there is another letter added. If you go from the top letter and follow a path to any other letter, each step you take adds a letter to a potential word. Each letter would have a boolean value, *isWord* that tells you if the collection of letters is a valid word. Graph tries (G-tries) are very similar. Each step you take adds a vertex to a graph, and has an attribute *isGraph* that tells you whether or not it is a valid graph. For the purposes of my research, the motifs of size 3 and size 4 will have this value set to true. Large

Figure 1.4: Network Motifs

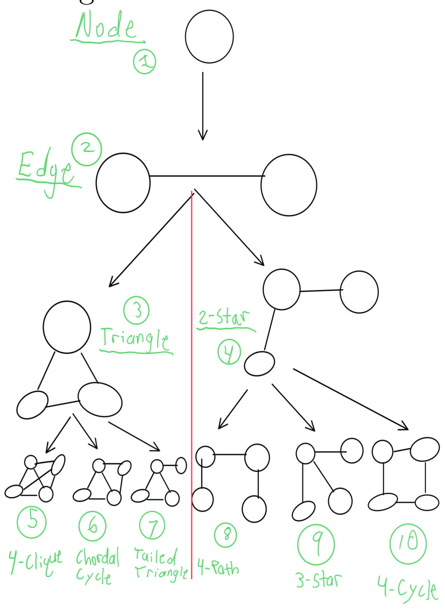
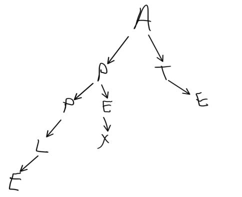


Figure 1.5: Trie Example



scale subgraph counting becomes much slower when computing subgraphs of size 4 or greater. Because of this, focusing on ”easy” subgraphs of size 3 and harder graphs of size 4 has shown to be a balanced approach that could easily scale to size n.

-Talk about value of shapes (and applications within real world networks)

1.4 Applications

biology structures

anonymous networks

comparing structures

random models for testing algorithms

comparing random models to real networks?

1.5 Algorithms