Introduction and Motivation: From tracing contamination in food supply chains [**?**] to gene interactions in C. elegans[**?**], network science has found a vast array of applications in the past two decades. This is largely a result of the generality of the network framework: myriad systems are readily adapted to such a description; interacting bodies (e.g. a city, person, or protein) form nodes in the network, the connections between them (e.g. via highways, Facebook friendships, or biological suppresion) create edges. Thus, one may usefully apply the same abstraction to study such disparate topics as the spread of opinions in a society and chemical reaction networks.
However, too often this merely leads to a recasting of the original problem. While this, in itself, can be useful, it fails to exploit the true power in such a formulation. Several obstacles prevent this full realization, a selection of which will be the focus of this proposal. Broadly speaking, the generality of the construct is in some sense its undoing: while numerous problems are addressable, they require many different types of analysis. Thus, a researcher in neural networks may use none of the same tools as a civil engineer designing tranporation networks. Certainly, the dynamics between proteins in a cell and drivers in rush-hour traffic may share almost no similarities; however, this should not prevent investigators in these fields from having similar analytical tools at their disposal.

Across many fields, the systems under investigation involve complex interactions in a large population. Consider the internet, thought to have over 4,000,000,000 URLs and continuing to evolve on the timescale of seconds [**?**]. Or the yearly human migration of millions across the world. The resulting data tends to be incredibly high dimensional, potentially involving a wide variety of seemingly disparate measures. Consider a researcher investigating the effect of immigration on foreign trade. A reasonable dataset would be composed of immigration figures to and from countries along with annual economic trade by industry, tracked over some time period. How should analysis proceeed? Initially, reasoning that increased immigration provides greater opportunities for commercial exchanges, one might try to correlate increases in immigration with rises in overall trade. But what about sector-specific gains? Such simple scrutiny might be misguiding: perhaps all of the immigrants to a given country were highly educated and growth was concentrated in technology firms. It could also be argued that immigration statistics should be correlated with trade figures several years in the future, as the effects of such movements are delayed as families situate themselves. Clearly, a well-defined toolbox of analytical methods for such problems would be helpful in guiding inquiries.

Thankfully, advances in data mining (or, more generally, dimensionality reduction) have provided a broad array of approaches to glean useful information from high-dimensional data. However, to date, such techniques have focused on data in the form of vectors; that is, they operate on a vector in $\mathbb{R}^n$, $n \gg 1$ and embed it in $\mathbb{R}^m$, $m \ll n$, where the m new dimensions capture the important details of the data. Unfortunately, when each data point takes the form of a graph, there is no clear way to "pre-embed" into $\mathbb{R}^n$ to make use of these existing algorithms. As we are increasingly able to track the evolution of dynamic networks, there exists a growing need to develop machinery capable of operating across networks. This proposal aims to address this gap in our abilities.

Proposed Research: Ideally, the large body of current dimensionality reduction techniques could be adapted to use on this new data type. Many existing methods employ some measure of distance between points in their formulation. For example, a common implementation of the popular diffusion maps algorithm requires a weight matrix $W_{ij} = e^{-\frac{\|x_i - x_j\|^2}{\epsilon}}$, in which x_i and x_j are members of the dataset. Thus, the first aim will be to devise a distance measure between to arbitrary graphs.

It turns out that even determining whether two graphs are equivalent (clearly a necessary ability when defining the similarity of two networks) is generally of **NP** complexity, and is called "the

graph isomorphism problem" (we consider two graphs to be equivalent if they are isomorphic). To this end, approximation algorithms exist, as do polynomial-time methods for certain special cases. Therefore, our research will initially focus on assessing the feasibility of these procedures for calculating the similarity of two graphs. This will require extending such algorithms to output not just a boolean true/false determination of isomorphism, but some metric that relates to certain differences in the two inputs.

In conjunction with this effort, we will also work towards improving graph kernels. Unlike the previous aim of adapting existing techonologies to the problem, this avenue of research considers completely new methods of dimensionality reduction specifically formulated to operate across graphs. Although a broad foundation of theory was established in a 1999 paper [**?**], there has been relatively little application of its findings to our particular focus. The methods that have appeared all seem to be based on random walks over the networks. These suffer from several limitations (outlined in [**?**]), and an alternative kernel based on deterministic network properties (perhaps subgraph densities) may alleviate these issues.

Finally, we must examine methods outside graph kernels. These would include the investigating the graph edit distance, which defines distances between graphs based on the additional number of nodes and edges that must be added before two graphs become isomorphic. Another approach is to align graphs so that "similar" vertices overlap. The problem admits for a great deal of creativity in its formulation, leading to a large variety of possible solutions. Indeed, even simple "common-sense based" approaches using subgraph densities or spectral information have proven effective in preliminary testing.

Applications: Such research would be relevant across numerous fields. Whenever the data is presented as a set of graphs, advances in these areas would provide an invaluable tool in the course of analysis. As mentioned, this sort of data would arise from a network's time evolution. A plausible application might be determining the cause of traffic jams in urban environments. We examine the transportation network (roads forming edges between city locations, weighted by the number of cars that occupy it at a given time), and use a data mining technique operating across the series of networks to identify the underlying variables that dictate traffic levels. It could also be used to analyze complicated models of disease spread to determine the appropriate measures to prevent an epidemic.

Apart from drawing conclusions from existing data sets, there is yet another application: coarse-grained analysis of complex network models. We've discussed how complicated these systems can be, it makes sense that simulating them might also present a challenge. The ability to examine the system from a macroscopic viewpoint using only a select few variables, instead of full system simulations, opens up a range of possibilities. From direct acceleration of runtimes via coarse projective integration, to the location of some macroscopic minimum using a coarse conjugate gradient method, our data mining techniques would provide the necessary low-dimensional description for such procedures. For instance, coarse minimization might be used to find the network configuration least susceptible to epidemics (??is that really a possible application??).

This project aims to address a major gap in current network analysis: general methods for extracting useful information from a set of graphs. Until this problem is addressed, researchers will continue to invent analytical techniques on a problem-specific basis, wasting time and potentially drawing innacurate conclusions. We should not allow this to continue.