Community Detection

In Large Networks

June Andrews

May 7, 2012

Thanks!

It goes without saying, these people have been inspiring forces of nature to work with:

- Mr. Len Kulbacki
- Coach Wilson
- Dr. James Sethian
- Patricia Kovatch
- Dr. Alex Vladimirsky
- Dr. John Hopcroft
- Dr. Steve Strogatz (thanks to Prof Rand for acting proxy :)
- Dr. Jon Kleinberg

Figure: Thanks to the NSF Graduate Fellowship Program!

Table of contents

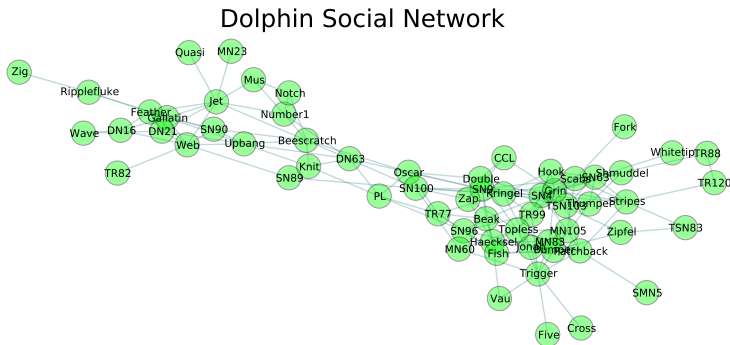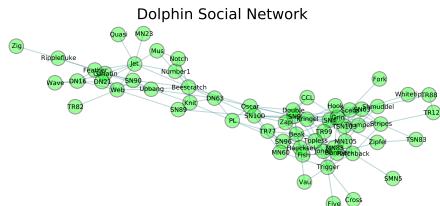# Dolphins

Dolphins form pods.

Figure: Nodes are dolphins. Edges are dolphins seen together.

Dolphins

If we know what the pods are:

- Who are the dolphins that interact between pods?
- How often do pods interact?
- Do the pods have a dominant leader?

Dolphin Social Network

Karate Club

People are members of a group.
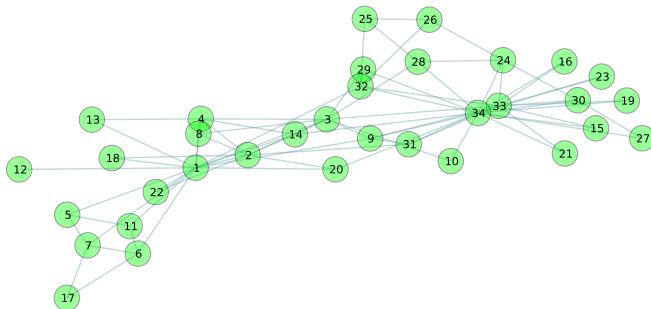
Karate Club Social Network

Figure: Nodes are students. Edges are students seen outside of class together.

Karate Club

If we know the groups people form:

- If something happens to split the social network, which people will do what?
- How many groups is someone a member of?
- Who are the influential members of a group?

Karate Club Social Network
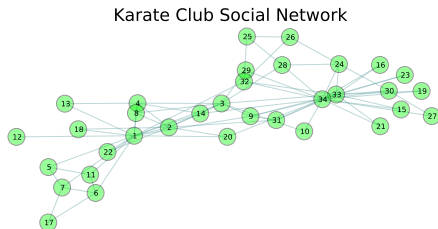
In General

JTODO insert figure of: Build Network $\rightarrow$ Find Communities $\rightarrow$ Analyze

How do you detect communities?

Science on Networks

Many applications have developed their own methods.

Application	Community Detection Method
Computation Distribution	Recursive Bi-section [Karypis & Kumar]
Statistical Mechanics	Belief Propogation [Hastings]
Storage of Large Matrices	Local Spectral Analysis [Andersen & Chung]
Taxonomy	Neighbor Joining [Saitou & Nei]
$\vdots$	$\vdots$

How do we compare them?

This Talk

- Create a way to understand Community Detection Methods.
- Create a method to handle large complex networks.
- Demonstrate power of new method on:
    1. Wikipedia Elections
    2. Physics Archive Citation Network

Brief Terminology

Definition (Internal Edges)

Edges between members of the same community.

JTODO include diagram of edges

Definition (External Edges)

Edges between members of different communities.

Previous Community Detection Methods

Pick a definition of a Community and then find it.

- (α, β) communities, every node in C is connected by at least β internal edges, every node outside of C is connected by at most α edges. [Mishra et al]
- Modularity, more internal edges than expected in a random graph. [Newman]
- Conductance, a step in a random walk will probabilistically stay within the community. [JTODO]
- Edge Betweenness, remove external edges to reveal communities. [Girvan & Newman]
- $\cdots$

Dection For Single Communities

Let's not define a community just yet.

Our approach:

- What are the desired characteristics of a community?
- Find a community with the best characteristics.

Characteristics of Single Communities

- INTERNAL DENSITY is density of edges within the community.
- EXTERNAL DENSITY is the density of edges leaving the community.
- SIZE of the community.
- DIAMETER of the community.
- AVERAGE SHORTEST PATH within the community.
- OUT DEGREE FRACTION, DEGREE DISTRIBUTION, $\cdots$

Representative Characteristics

All of the listed characteristics can be bounded by INTERNAL DENSITY, EXTERNAL DENSITY, and SIZE.

The 3D space of Communities

Given a metric M that evaluates communities in the $(I(C), E(C), |C|)$ space, let us find
If networks were in a continuous space, we could use gradients to maximize metric.

Conductance

Level sets of Conductance.

New Metric for a Single Community

Level sets of

Algorithmcs in the IE space

Running algorithms in the IE space.

Characteristics of Sets of Communities

Many characteristics, again defined by internal density, external density, and size of set.

The 3D space of Communities

Consider level sets of how a metric divides the space.
If continuous use gradient to maximize metric.

Modularity

Level sets of Modularity.

New Metric for a Single Community

Level sets of

Algorithm for maximizing new metric

Conjecture, once you decrease int density, never go back
set $a = 1$ and increment b, and c.
Use the Louvain algorithm for maximizing the metric at each stage.

Algorithmcs in the IES space

Running algorithms in the IES space.

Future

Networks are getting big.
Before we asked the question of, is this a good set of communities.
Now we need to ask does this node and set of nodes belong to a community?
Why? No communication in parallelization.

Characteristics

The two characteristics between a node and a set of nodes are χ_e and χ_p.

Community

How do we know we have a community?
Closure.

Algorithm

- Find the seeds
- Expand the seeds

Seeds

Pick a set of nodes we know must belong to the same community.
Find very dense sets of nodes, presume begin with a clique, and expand
into the local neighborhood.
If seeds are distance 2 apart will be in the same community, if the
community is large.

Expansion

Correctness of Expansion

Probability of Correctness.

Algorithms

Checkpoint

We have new methods.

We'll leave comparisons to known methods to the thesis.

We'll do a sanity check against known communities and then present results from applications that could not have been done before.

Test on Dolphins

Test on Karate

Physics Archive

The more communities a paper is immediately popular in the more citations that paper will get.
Correlated with cross community journals - confirm.

Wikipedia Voting

JTODO include picture of what it is

Communities Predicting a User's Vote

Communities are sets of user with similar voting patterns. There are ≈ 600 communities covering *JTODO*% of the nodes.

Vote Prediction

Given the communities a user is in, we can predict a user's vote with 86% accuracy.

This is close to Kleinberg et al's work of 90% accuracy.

Communities Predicting an Election's Outcome.

Users campaign to sets of communities, but not everyone from a community votes. If they did:

Election Prediction

Given the Communities in Wiki Voting Similiarities, 14% of election results would be over turned.