Community Detection

In Large Networks

June Andrews

Cornell University

May 8, 2012

Thanks!

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

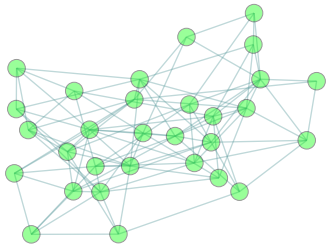
- Mr. Len Kulbacki
- Coach Wilson
- Dr. James Sethian
- Patricia Kovatch
- Dr. Alex Vladimirsky
- Dr. John Hopcroft
- Dr. Steve Strogatz (thanks to Prof Rand for acting proxy :)
- Dr. Jon Kleinberg

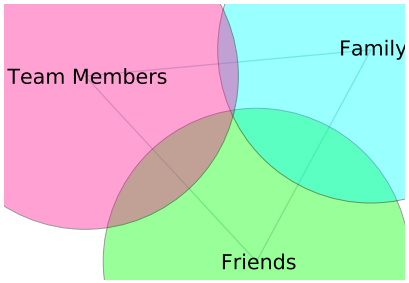
Figure: Thanks to the NSF Graduate Fellowship Program!

Table of contents

Applications In General

Build Network → Find Communities → Analyze





Who are the leaders?
How do groups interact?

How do you detect communities?

Dolphins

Dolphins form pods.

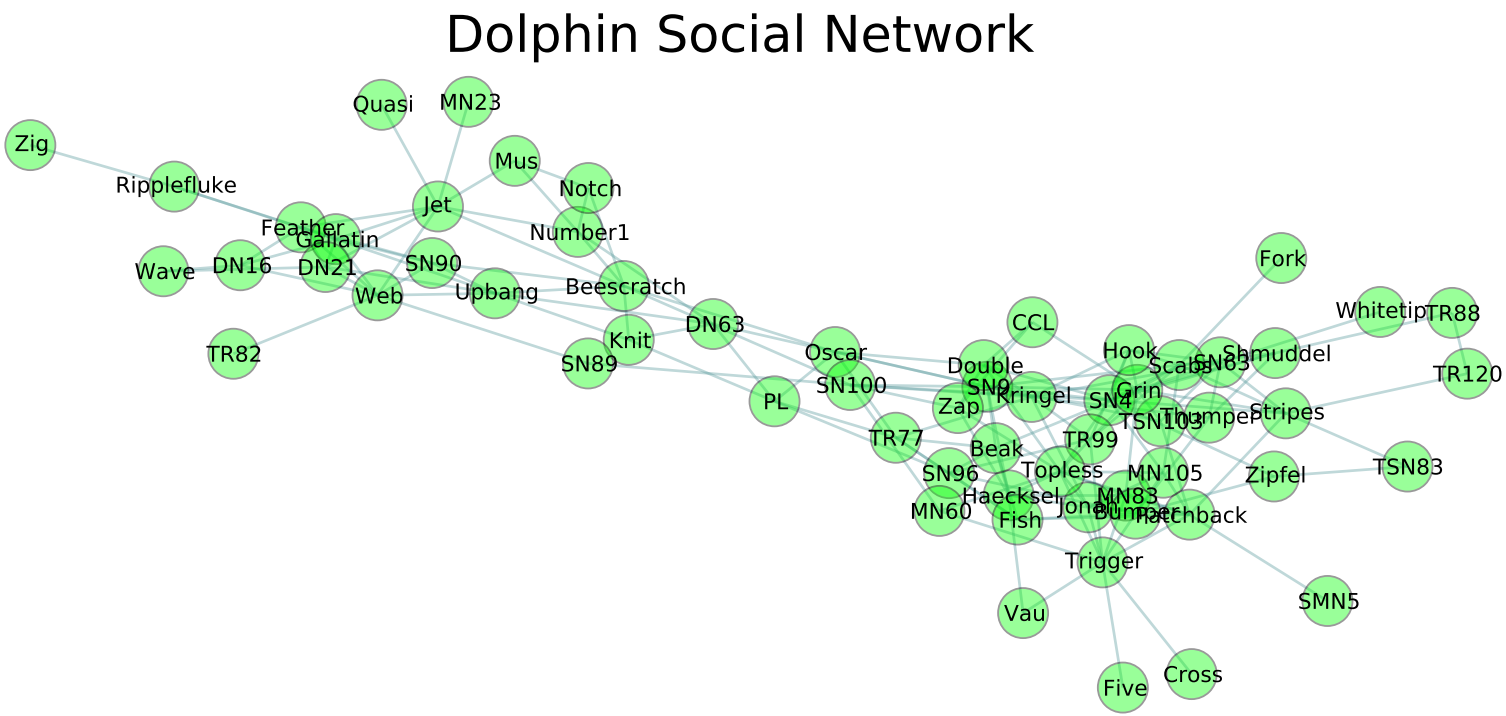
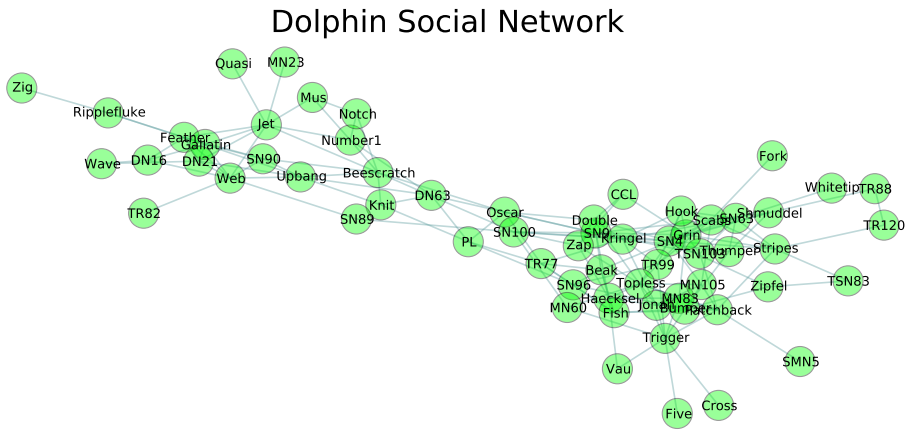


Figure: Nodes are dolphins. Edges are dolphins seen together.

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?



Karate Club

People are members of a group.

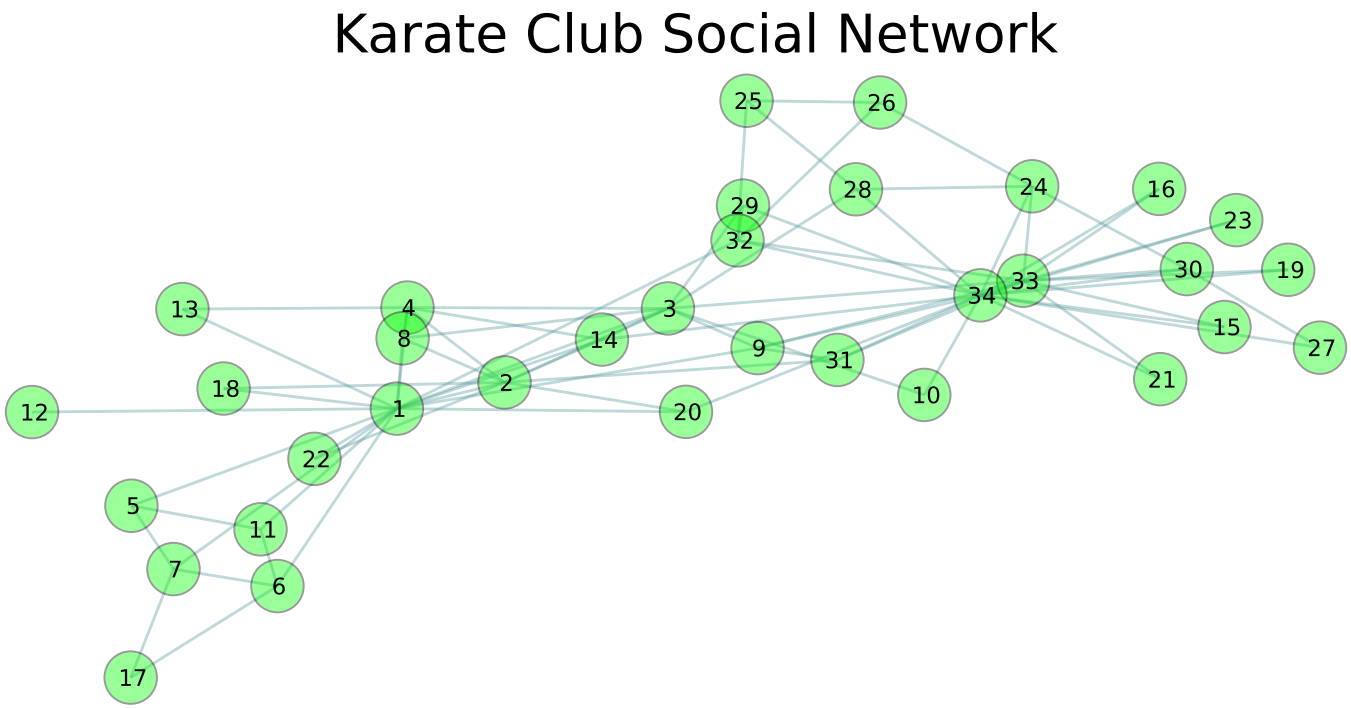
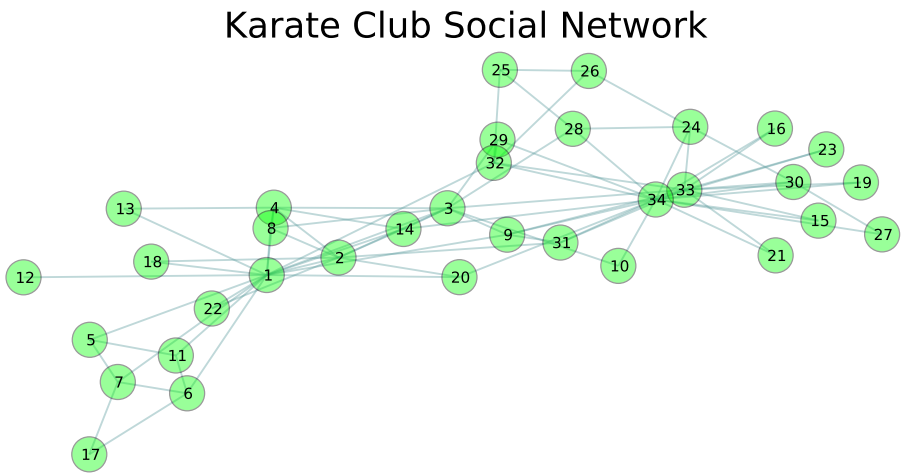


Figure: Nodes are students. Edges are students seen outside of class together.

Karate Club

If we know the groups people form:

- If students disagree, who will disagree with who?
- How many groups is someone a member of?
- Who are the influential members of a group?



Science on Networks

Many fields of applications have developed their own methods.

Application	Community Detection Method
Computation Distribution	Recursive Bi-section [Karypis & Kumar]
Statistical Mechanics	Belief Propogation [Hastings]
PageRank	Local Spectral Analysis [Andersen & Chung]
Taxonomy	Neighbor Joining [Saitou & Nei]
⋮	⋮

How do we compare them?

Experimental comparisons have been made for certain networks.

Size of Networks

Applications are increasing in size as they become available.

Network	Number of Nodes	Date Available
Karate Club	34	1970s
Astrophysics Co-authors	19k	1999
Twitter	20 million	2009
⋮	⋮	⋮
Facebook	1 Billion	???

We need a parallel method that scales.

Pioneer parallel methods are being developed.

Complexity of Communities, always increasing

Subcommunities of a Physics Archive Community

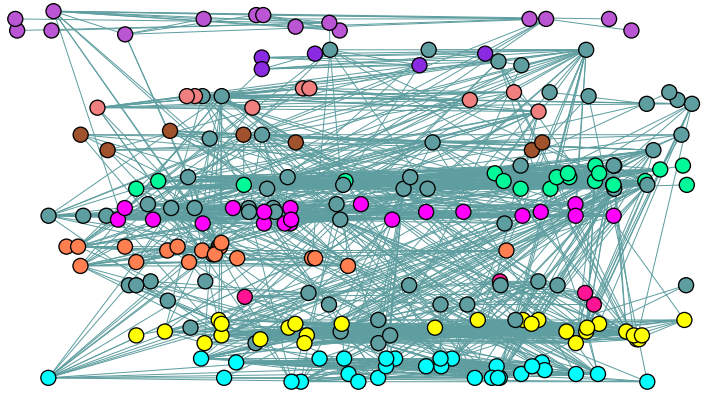


Figure: The umbrella community is results on computing *flow equations*. The subcommunities are different approaches to calculating the equations.

We need a method for weighted networks and overlapping communities.

Only predictably overlapping communities have been found.

This Talk

- Framework to understand Community Detection Methods.
- Parallel method for large complex networks.
- Demonstrate power of new method on:
  - Wikipedia Elections
  - Physics Archive Citation Network

Brief Terminology

Definition (Internal Edges)

Edges between members of the same community.

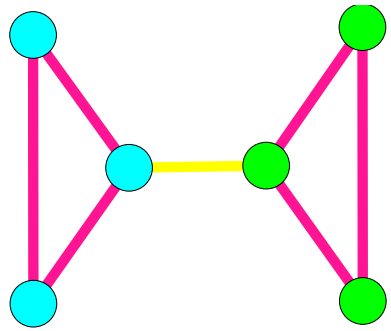


Figure: Internal Edges are in Pink, External Edges are in Yellow.

Definition (External Edges)

Edges between members of different communities.

Previous Community Detection Methods

There is no universal definition of a community. Methods pick a definition of a Community and then find those communities.

- (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, probability a step in a random walk will leave the community. [used in Andersen & Chung]
- Edge Betweenness, remove external edges to reveal communities. [Girvan & Newman]
- $\cdots$ many more

Characteristics of Single Communities

Characteristic	Desired Value
Internal Density	large number of internal edges
External Density	small number of external edges
Size	application specific
Diameter	short
Average Shortest Path	short
Out Degree Fraction	small
Degree Distribution	application specific
⋮	

Representative Characteristics

All of the listed characteristics of community C can be bounded by Internal Density, $I(C)$, External Density, $E(C)$, and Size, k.

Metrics

Metrics collapse all characteristics of a community, C, down to a real number.

Definition (Metric for a Single Community)

$$M : \{C \in Communities\} \rightarrow \mathbb{R}$$

Some metrics are only functions of internal density, external density, and size of a community. We assume continuity.

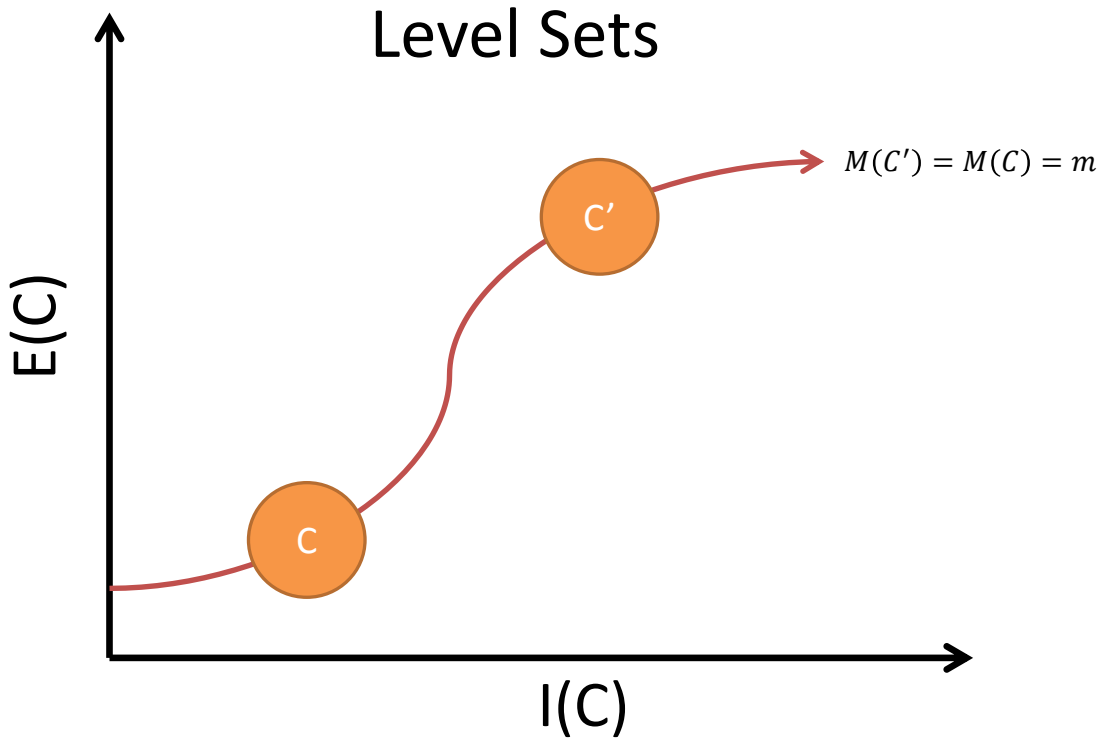
Definition (Metric for a Single Community)

$$M : (I(C), E(C), k) \rightarrow \mathbb{R}$$

We can visualize how these metrics collapse the $3D$ space describing these communities to the real numbers. We cover the dimension of size in the paper, but only consider internal density and external density here.

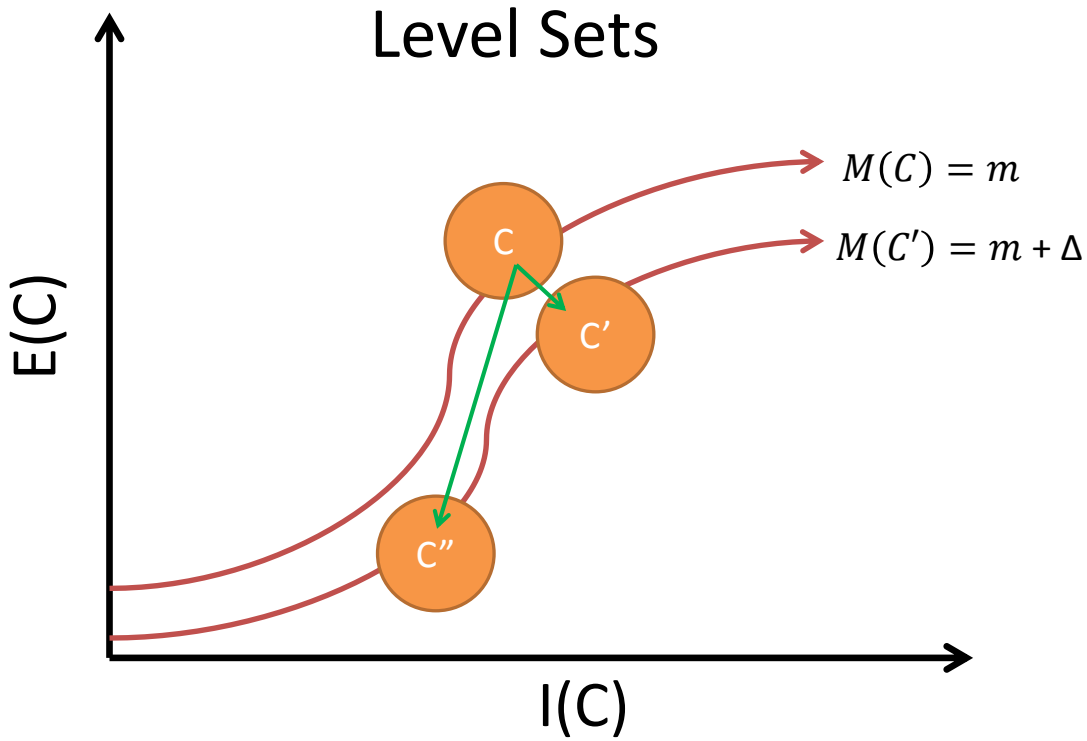
The 3D space of Communities

Consider the set of $\{(I(C), E(C)) \in [0,1]x[0,1]\}$ values, such that $M(C) = m$. This forms a level sets. We can categorize the (I, E) space with level sets.



The 3D space of Communities

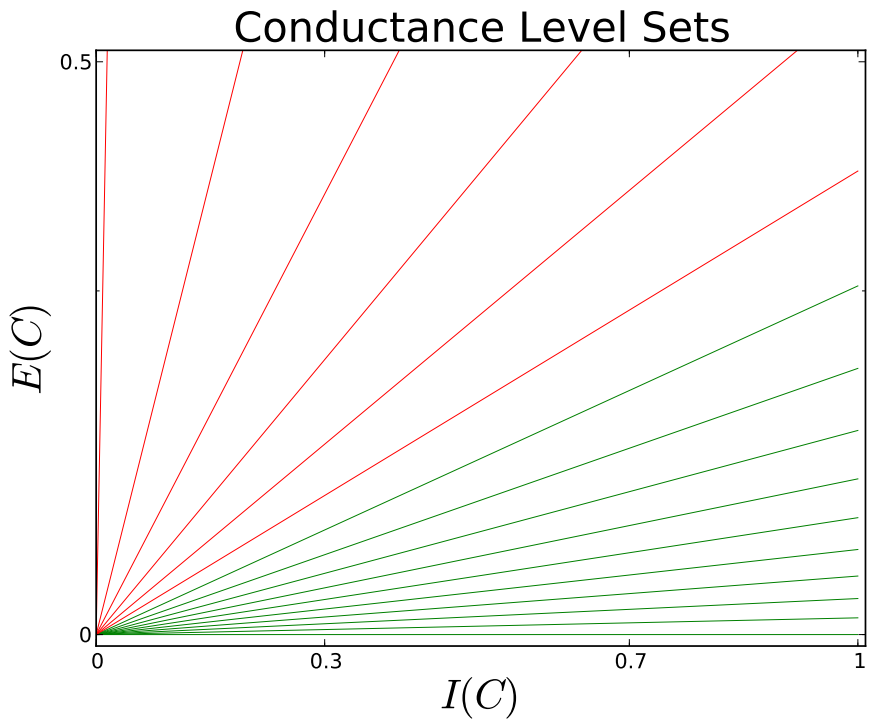
Consider the set of $\{(I(C), E(C)) \in [0,1]x[0,1]\}$ values, such that $M(C) = m$. This forms a level sets. We can categorize the (I, E) space with level sets.



Conductance

Probability a random step leaves the community.

$$\textsc{Conductance}(C) = \frac{(1-k)E(C)}{kI(C) + (1-k)E(C)}$$



Poor $I(C)$ value ↗

↖ Good $I(C)$

Conductance

A series of communities created by including nodes that improve conductance.

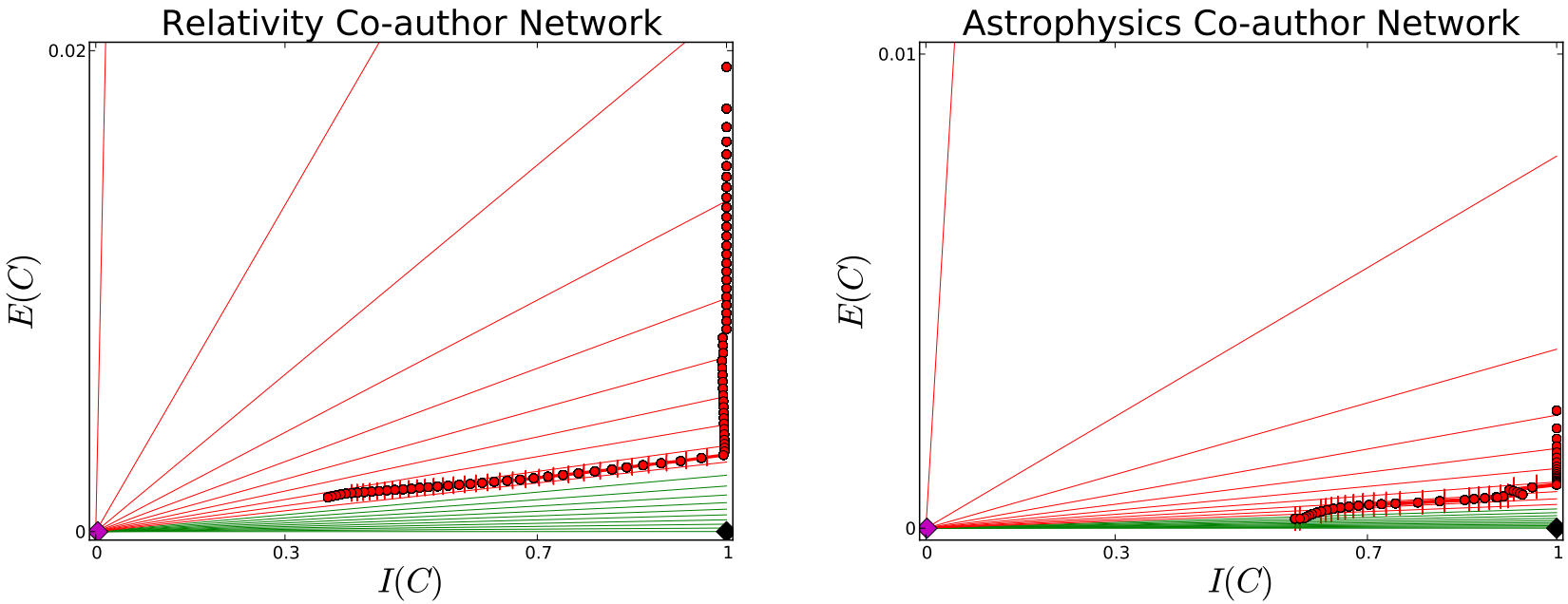


Figure: The co-authors in relativity ($\approx 3k$ nodes) and astrophysics ($\approx 18k$ nodes).

Makes small improvements to external density at great cost to internal density.

Modularity of Single Community

Difference in present versus expected edges in a random graph.

$$\textsc{Modularity}(C) = pI(C) - (pI(C) + qE(C))^2.$$

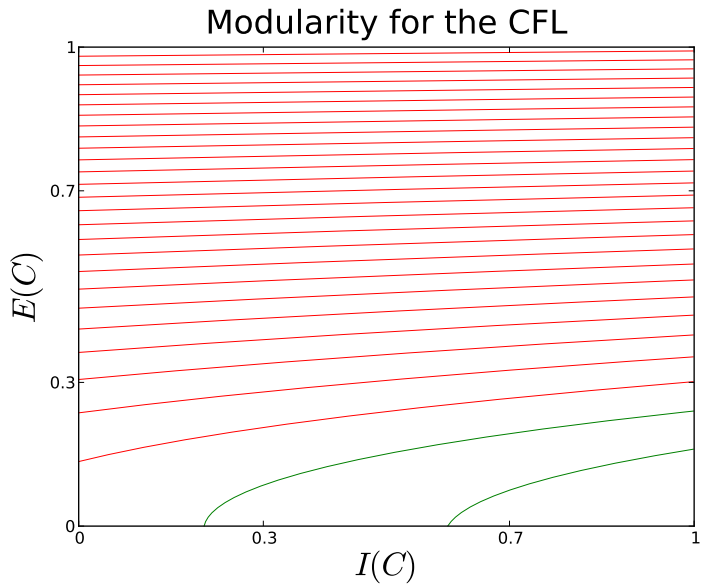


Figure: Level Sets have the same affect as Conductance, until external density is low enough and the level sets prevent further loss in internal density.

Modularity of Single Community

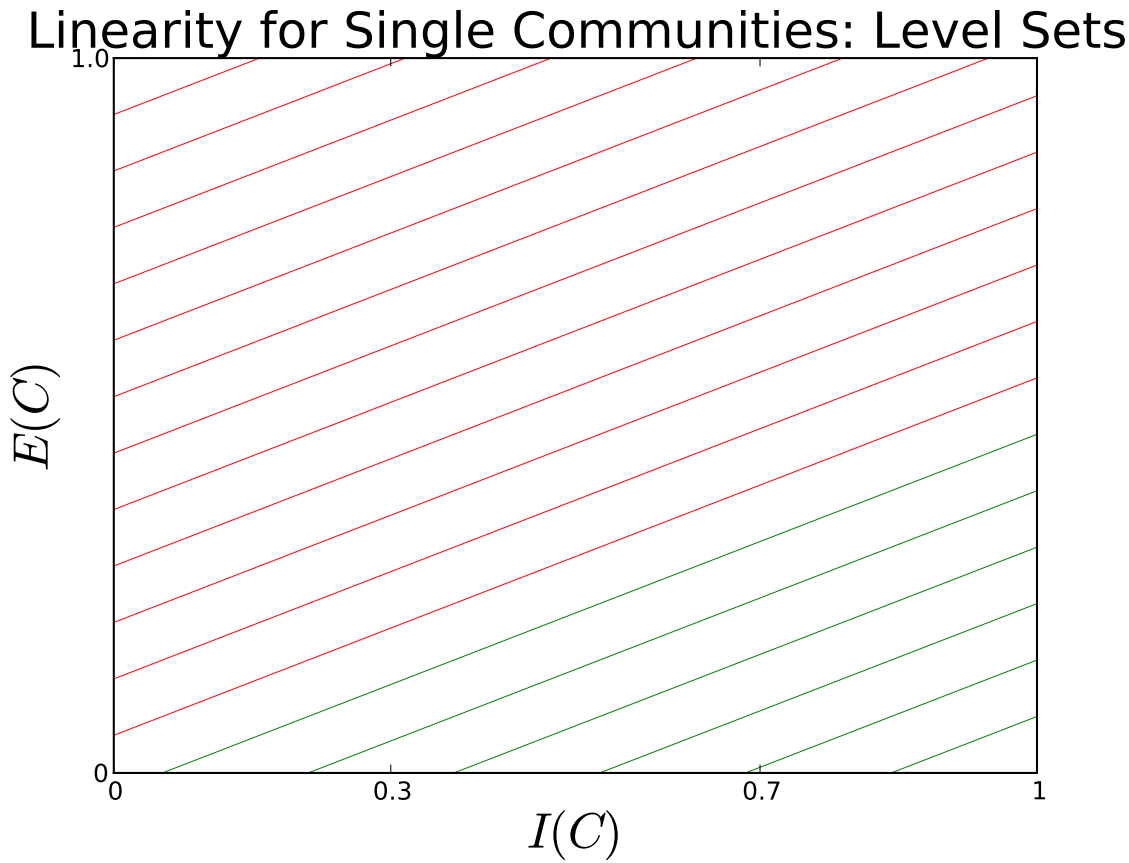
JTODO include picture of modularity steps for a single community. Optimizes external density and then internal density. Optimizations of external density join together communities. This is what causes the resolution limit in modularity.

More Single Community Metrics

Metric	definition	Results in
Conductance	probability a random step leaves community	≈disconnected components
Edges Cut	number edges connecting the community and graph	≈disconnected components
Expansion	average number of edges per node leaving the community	≈disconnected components
Internal Density of Newman	number of edges within a community	cliques

New Metric for a Single Community

$$\textsc{Linear}(C) = aI(C) - bE(C)$$



New Metric for a Single Community

JTODO improve

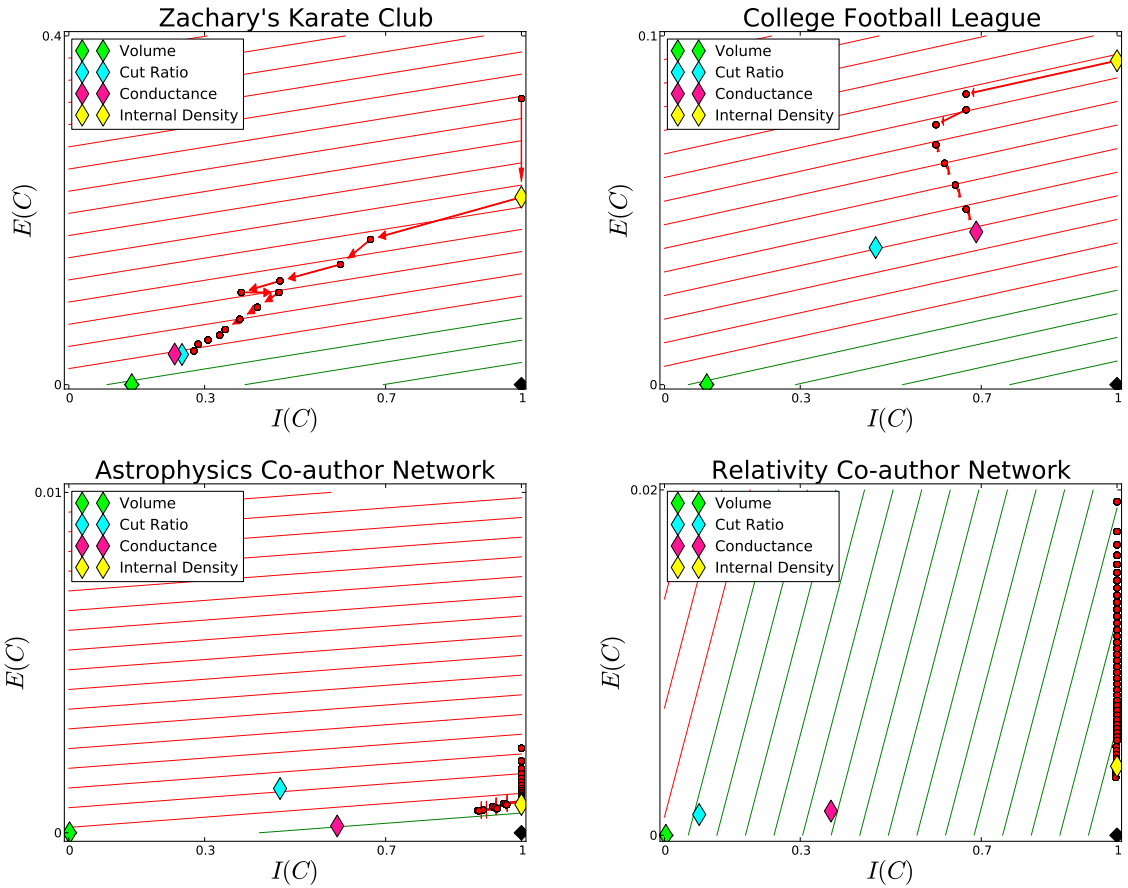


Figure: Linear is consistently returns strong communities.

Characteristics of Sets of Communities

Same characteristics of single communities, just adapted.

Characteristic	Represents
Internal Density	How close communities are to cliques.
External Density	How many edges are not within a community.
Size	Number of communities.
Average Diameter	
Average ...	

Representative Characteristics

Listed characteristics of communities, $S = \{C_1, C_2, \dots\}$, can be bounded by Internal Density $I(S)$, External Density $E(S)$, and Size $|S|$.

New Metric for a Sets of Communities

Linear(S)

For constants, $a, b, c \in (0, 1)$ and set of communities $S = \{C_1, C_2, \dots\}$:

$$\text{Linear}(S) = aI(S) - bE(S) - c|S|$$

Definition (Internal Density of a Set of Communities)

$$I(S) = \frac{\sum_{C \in S} \text{Number of edges in } C}{\sum_{C \in S} \frac{1}{2}|C|(|C| - 1)}$$

Definition (External Density of a Set of Communities)

$$E(S) = \frac{\text{Number of Edges not in some } C}{\text{Number of Edges in the Graph}}$$

Tools For Building an Algorithm

Can use the fast Louvain heuristic algorithm to find a partition that maximizes $\textsc{Linear}(S)$.

Conjecture

A greedy algorithm can (*almost always*) only decrease internal density.

The known exception occurs when a comparatively large clique is involved, very rare.

Greedy Algorithm for Linear

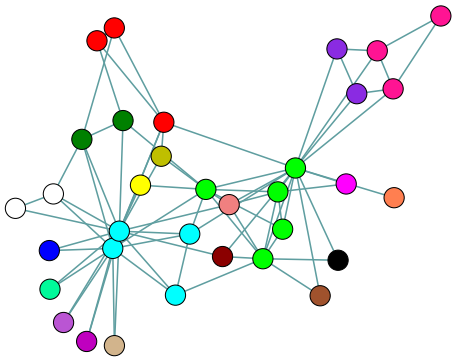
Find Cliques
$(a, b, c) = (1, 0, 0)$

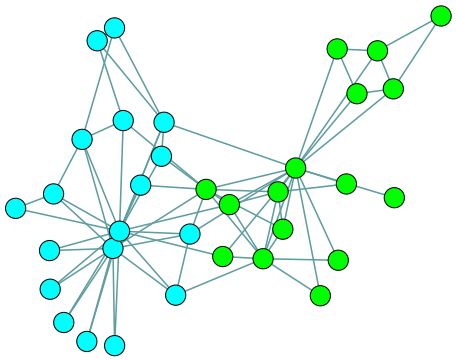
$\rightarrow$

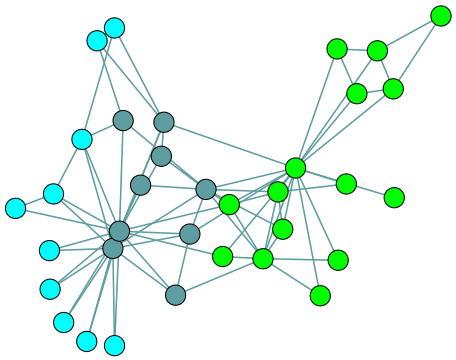
Find Partition
$(a, b, c) = (1, \Delta, \Delta)$

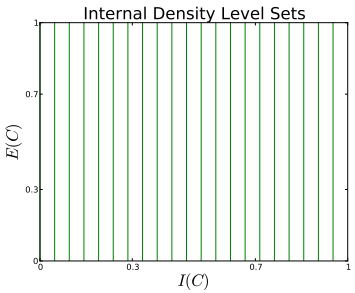
$\rightarrow$

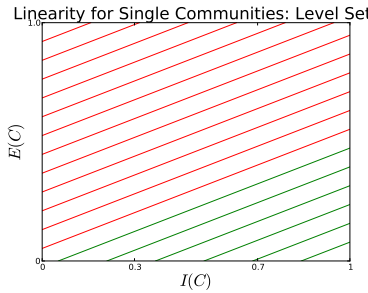
Expand Partitions
$(a, b, c) = (1, \Delta, \Delta)$











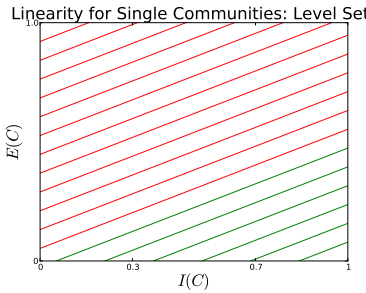


Table: Communities are indicated by node color. Nodes in multiple communities are grey.

Algorithm for Linear in the IE plane

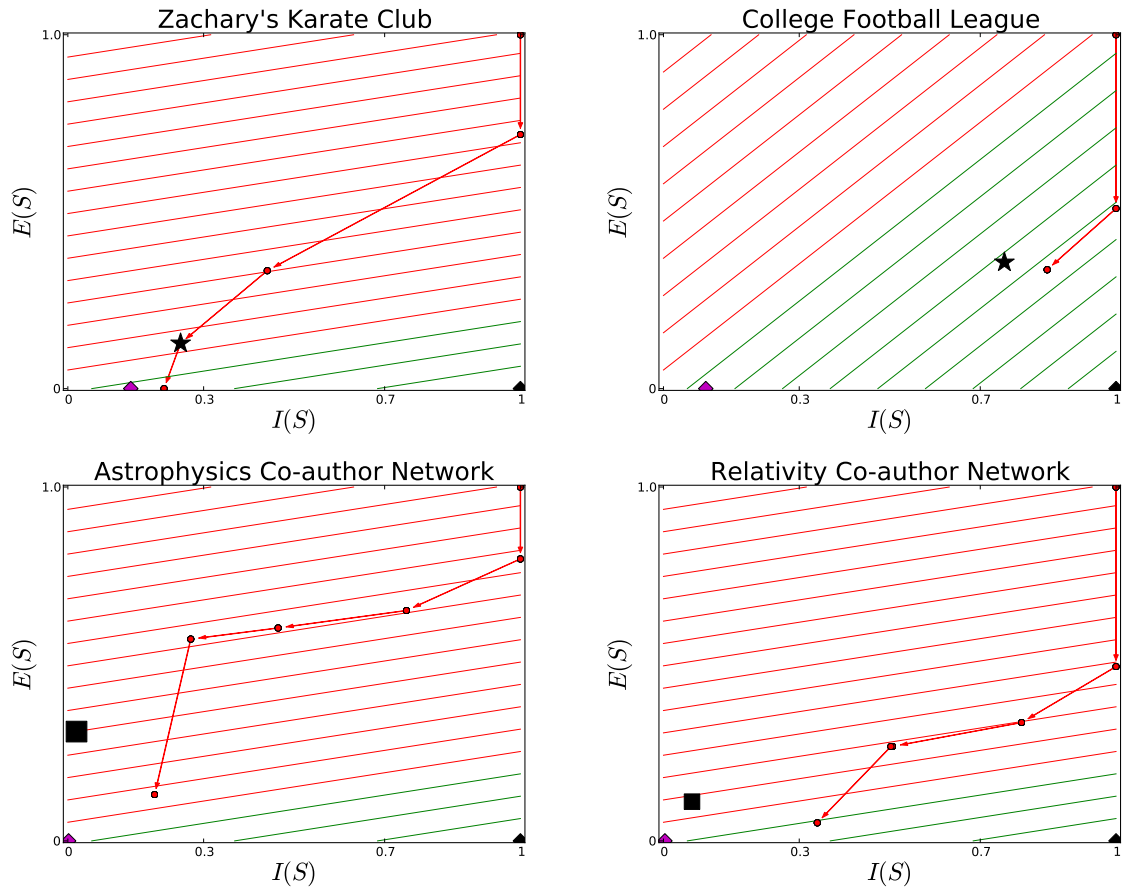


Figure: Black stars and squares are the sets of communities found by modularity

Future

Small Network Paradigm

Is C_1 a good community within a set of communities $\{C_1, C_2, \dots\}$?

Does not parallelize.

Large Network Paradigm

Given a node n and the set of nodes $X = \{n_1, n_2, \dots\}$, do they belong to the same community C?

Easily parallelized.

Large Network Paradigm has been used by [Mishra et al] and [Wang & Hopcroft]

Characteristics

Definition (χ_e)

The number of edges a node has to members of a set of nodes, X, is characteristic χ_e.

$$\chi_e(n, X) = |\text{Neighbors}(n) \cap X|$$

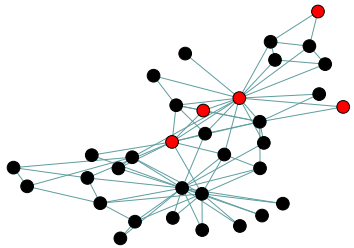
Definition (χ_p)

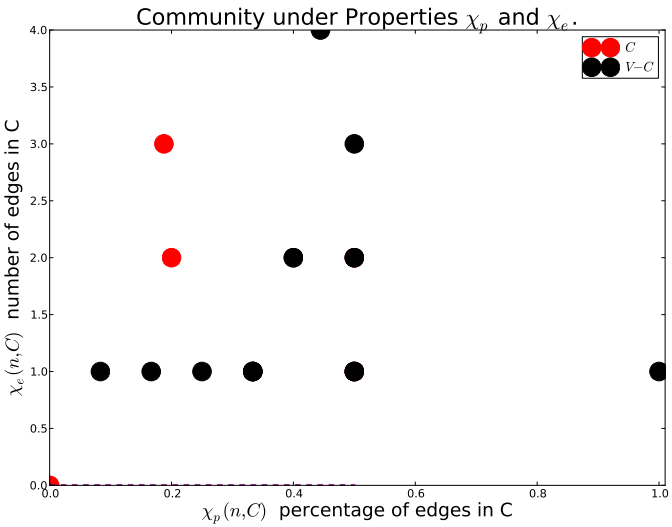
The proportion of total edges a node has to members of a set of nodes is characteristic χ_p.

$$\chi_p(n, X) = \frac{|\text{Neighbors}(n) \cap X|}{\text{degree}(n)}$$

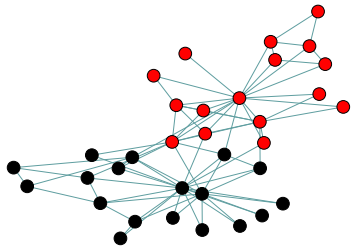
If $\chi_p(n, X)$ **OR** $\chi_e(n, X)$ is high, then $n \cup X$ belong to a community.

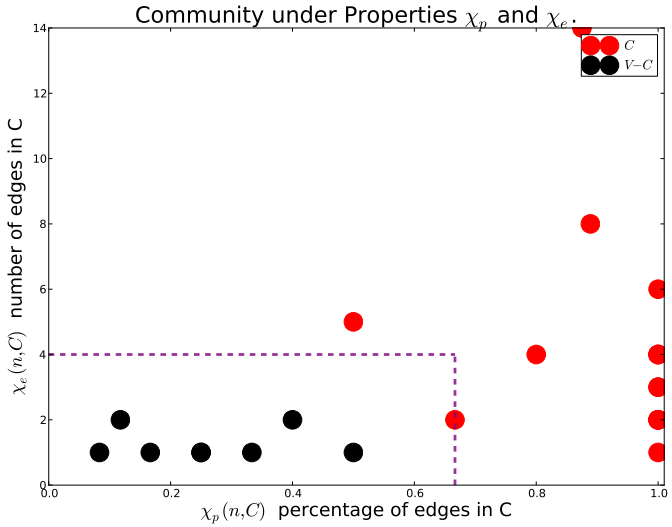
Open Community





Closed Community





Seeds

To begin need a set of nodes, X, that belong to the same community.

Definition (Seed)

A set of nodes that belong to the same community.

Strongest seeds are near cliques. Easy to find. Can be done in parallel.

Expansion

Given the set of seeds, we can expand them in parallel to find closed communities.

Communities= []
for $X \in$ Seeds **do**
  while there exists $n \notin X$ with a high $\chi_e(n, X)$ **OR** $\chi_p(n, X)$ **do**
    $X \leftarrow X \cup \{n\}$
  end while
  Add X to Communities
end for

To optimize select n with the maximum $\chi_e(n, X)$ **OR** $\chi_p(n, X)$.
Variations exist to find larger communities.

Correctness of Expansion

In the thesis we find closed form solutions for:

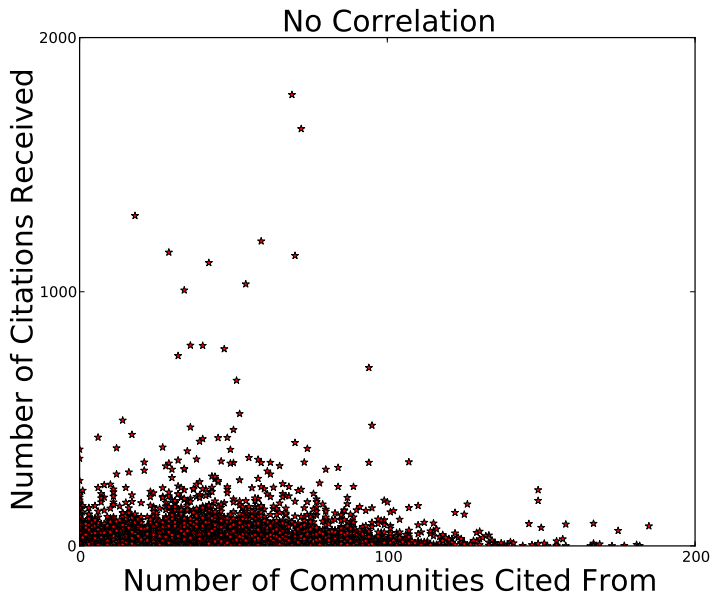
$$P(n \cup X \subset C \mid n \text{ has max } \chi_e(n, X))$$
$$P(n \cup X \subset C \mid \chi_p(n, X))$$

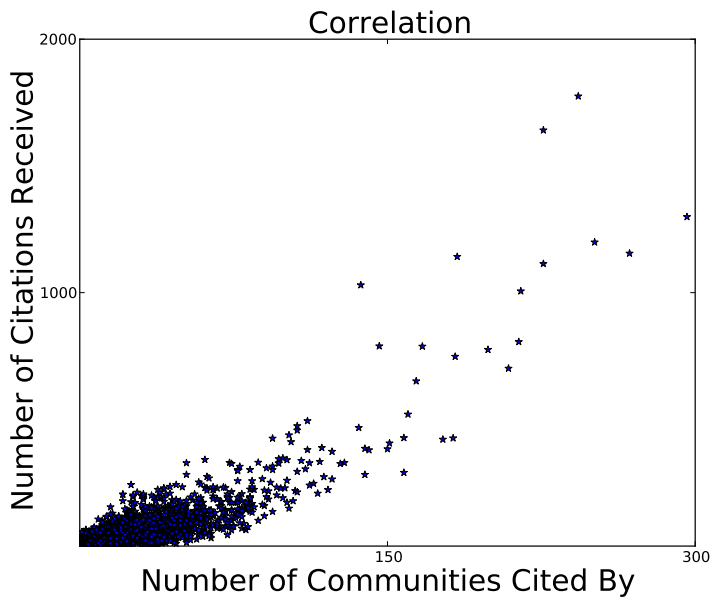
Select	Binomial Random Graph	Power Law Graph
node n has max$\{\chi_e\}$	$> 94\%$	$> 90\%$
$\chi_p(n, X) = \frac{1}{2}$		

Table: Probability of Correctly Expanding a seed under unfavorable conditions.

Physics Archive

We know the number of citations a paper makes is not related to the number of citations a paper receives[]. Same is true for communities.





But, the more communities a paper is cited by, is correlated with a paper's success.

Correlated with cross community journals - confirm.

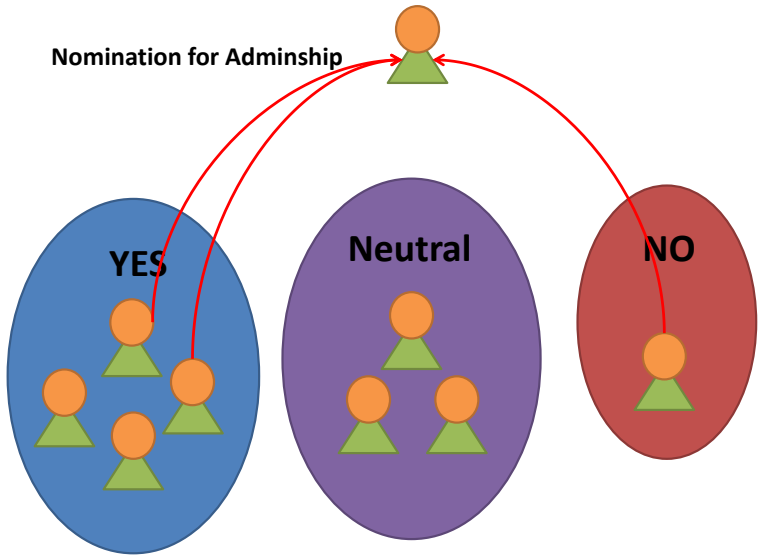
Physics Archive

JTODO include plot of communities evolving, we can now pick out the nodes corresponding to spanners, interpretors, etc

Physics Archive

JTODO include diagram of number of communities a node belongs to follows a power law distribution

Wikipedia Voting, 2794 elections



Network: 6287 users that vote $114k$ times in 2794 elections (*not everyone votes*). There are ≈ 624 communities, with an average size of 24 users. Communities are sets of user with similar voting patterns, in sign and frequency. Users not in communities have erratic voting patterns. The number of communities a user is in, follows a power law distribution.

Communities Predicting a User's Vote

To predict how a user voted in an election:

- Remove the user's vote.
- Calculate how the communities the user is in, voted.
- Predict the user's vote with how a majority of the communities voted.

Accuracy	Percentage of Votes Predicted
83%	60%
86%	13%
95%	4%

Table: The affect of smaller communities

[Leskovec, Huttenlocher, & Kleinberg] attain 90% accuracy with a model for vote predictions.

Communities Predicting an Election's Outcome.

We double the number of votes, by including votes with 90% accuracy.

Election Prediction

If twice as many people voted, 7% of elections would be overturned.

Did Vote	If More Users Voted	Type
{Yes: 158, No: 60}	{Yes: 498, No: 82}	False Negative
{Yes: 14, No: 6}	{Yes: 109, No: 25}	False Negative
{Yes: 45, No:7}	{Yes: 93, No: 38}	False Positive

Table: Sample Changes in the Vote

Application

Use the communities to detect if a non-representative group is voting and alert Wikipedia Bureaucrats, who can overturn the election.