Network Properties Revealed through Matrix Functions

Qun Su

11-8-2017

Outline

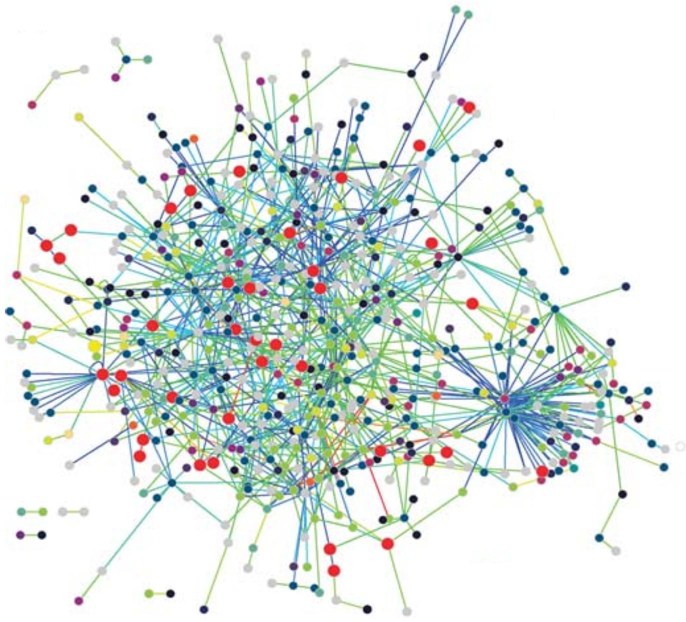
- Motivation and background
- Measuring network with matrix exponential
- New set of measures: matrix resolvent
- Relation with graph Laplacian and spectral clustering
- Resolvent vs. exponential
- Experiment
- Conclusion

Outline

- Motivation and background
- Measuring network with matrix exponential
- New set of measures: matrix resolvent
- Relation with graph Laplacian and spectral clustering
- Resolvent vs. exponential
- Experiment
- Conclusion

Motivation

- Analysis and characterization of networks offers valuable guidance in many areas
  - Cell biology: gene-protein connections
  - Brian: interconnection in neurological regions
  - Epidemiology: epidemical contact of people
  - Zoology: social interaction among animals
  - Energy: electricity transport network
  - Telecommunication
  - WWW
  - Movie database: costarring



Protein interaction network of Treponema pallidum

- Networks are typically complex → the paper aim to describe complex networks with some simple quantities
- Network: undirected, unweighted graph with N nodes

Node centrality

- Measuring networks with the concept of centrality was proposed decades ago.
- Most intuitive method: Freeman centrality (Freeman, 1979), or degree

$$\deg_i := \sum_{k=1}^{N} a_{ik} = (A\mathbf{e})_i$$

  - Counts the number of edges connecting to node *i*
  - *e* is a vector with all elements being 1
  - *A* is the adjacency matrix of the network
- Katz centrality (Katz, 1953), an extension of Freeman centrality

$$k_i := \sum_{j=1}^{N} \sum_{k=0}^{\infty} \alpha^k (A)_{ij}^k = \left(\left((I - \alpha A)^{-1} - I\right)\mathbf{e}\right)_i$$

  - *I* = *N*-dimensional identity matrix
  - *α* is a fixed parameter. Its upper bound is the inverse of *A*'s largest eigenvalue

Node centrality

- Eigenvector centrality (Bonacich, 1987), for weighted networks

$$b_i := \frac{1}{\lambda_1} \sum_{j=1}^{N} a_{ij} b_j = \left(\frac{1}{\lambda_1} A\mathbf{f} \right)_i$$

  - λ_1 – Perron-Frobenius eigenvalue of A
  - f – Perron-Frobenius eigenvector of A

Walk vs. path

- A path between node *i* and node *j* (*i* and *j* are distinct) is an ordered list of distinct nodes $i, k_1, k_2, \ldots, k_{n-1}, j$, in which successive nodes are connected.
- Walk between node *i* and node *j* is an ordered list of nodes $i, k_1, k_2, \ldots, k_{n-1}, j$, in which successive nodes are connected.
  - The start and end of a walk may be the same ($i = j$)
  - Nodes may be revisited ($k_1, k_2, \ldots, k_{n-1}$ are not necessarily distinct)
- Lemma 1.1: The quantity $(A^n)_{ij}$ counts the number of different walks ($i \neq j$) or closed walks ($i = j$) of length n between nodes *i* and *j*.

Outline

- Motivation and background
- Measuring network with matrix exponential
- New set of measures: matrix resolvent
- Relation with graph Laplacian and spectral clustering
- Resolvent vs. exponential
- Experiment
- Conclusion

Matrix exponential: Centrality

- Degree of node *i* can be alternatively interpreted by number of closed walks of length 2 from *i*

$$(A^2)_{ii} = \sum_{k=1}^{N} a_{ik}a_{ki} = \deg_i$$

- Consider lemma 1.1, $(A^n)_{ii}$ gives the number of closed walks involving node *i*, which reflects how *i* is connected to the network
- Intuitively, we write the centrality of node *i* as

$$(A^2)_{ii} + (A^3)_{ii} + (A^4)_{ii} + \dots$$

- Note that longer walks are less efficient than shorter walks. Hence, we need to add a weight factor to each term such that longer walks contribute less to centrality

Matrix exponential: Centrality

- Let the weight factor be *1/(n!)* with n being the walk length and add an constant bias *(I + A)*

Arbitrary constant

Weighted sum of closed walks with all possible length

$$\left(\boxed{I + A} + \underbrace{\frac{A^2}{2!} + \frac{A^3}{3!} + \cdots + \frac{A^k}{k!} + \cdots}_{} \right)_{ii}$$

- Can be rewrite to

$$(\exp(A))_{ii}$$

Matrix exponential: Communicability

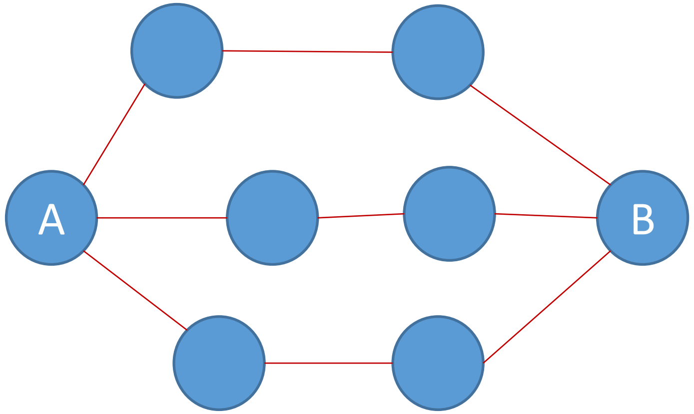
- Communicability quantifies the easiness for a piece of information to pass from node *i* to node *j (i* and *j* are distinct*)*
- A reasonable expression for communicability is sum over all walks that connect *i* and *j*.

$$(A^2)_{ij} + (A^3)_{ij} + (A^4)_{ij} + \ldots$$

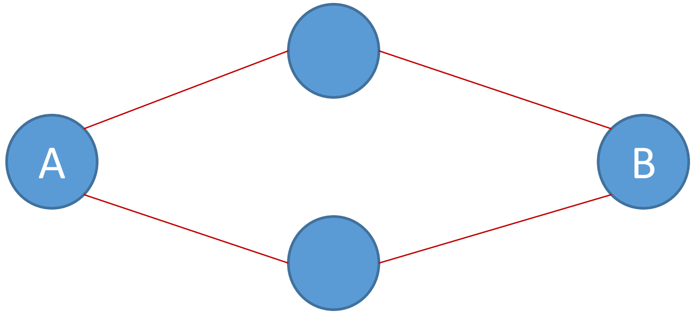
- Again, longer walks through *i* and *j* are penalized for not being efficient. If still use *1/(n!)* as the weight and *(I + A)* as bias, we have

$$(\exp(A))_{ij}$$

1/3! + 1/3! + 1/3! = 1/2



1/2! + 1/2! = 1/2



UNIVERSITY OF MINNESOTA
Driven to Discover

Matrix exponential: Betweenness

- Betweenness quantifies the importance of a particular node for information flow within the network. Alternatively, it quantifies the change of overall communicability of the network if a particular node is removed.
- Denote the node to be removed as *r*, and let *E(r)* be a matrix whose components are nonzero only in (1) row and column *r*, AND (2) *A* has 1 in that position. Then the change in communicability per pair of nodes (other than *r*) is

$$\frac{1}{(N-1)^2-(N-1)}\sum\sum_{i\neq j, i\neq r, j\neq r}\frac{(\exp(A)_{ij}-\exp(A-E(r))_{ij})}{\exp(A)_{ij}}$$

  - $N \geq 3$
  - Number of terms in the summation is $(N-1)^2-(N-1)$
- Up to now, we developed a methodology of using matrix exponential as a measure of a network in three aspects: centrality, communicability, and betweenness

Outline

- Motivation and background
- Measuring network with matrix exponential
- New set of measures: matrix resolvent
- Relation with graph Laplacian and spectral clustering
- Resolvent vs. exponential
- Experiment
- Conclusion

General rule of forming new network measures

- Assumption: the graph is connected, with N nodes.
  - There must be at least one walk of length less than N between two nodes.
- Generally, to propose a new set of network measures, we write centrality in the form:

$$\sum_{n=1}^{\infty} c_n A^n$$

with $c_n \geq 0$ being the weight to scale the number of walks of length n.

- $c_n = 1/(n!)$ is not the only way of scaling.

General rule of forming new network measures

- A complete set of network measures should have the following form:
  - f-centrality: given by $f(A)_{ii}$, or in terms of the spectrum of A, $\sum_{k=1}^{N} f(\lambda_k) {x_i^{[k]}}^2$. $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_N$ are the eigenvalues of A corresponding to eigenvectors $x^{[1]}, x^{[2]}, \ldots, x^{[N]}$.
  - f-communicability: given by $f(A)_{ij}$ (i and j are distinct), or $\sum_{k=1}^{N} f(\lambda_k) x_i^{[k]} x_j^{[k]}$
  - Betweenness: given by

$$\frac{1}{(N-1)^2-(N-1)} \sum \sum_{i \neq j, i \neq r, j \neq r} \frac{(f(A)_{ij} - f(A-E(r))_{ij})}{f(A)_{ij}}$$

- From the factorial weight, a new f should
  - Penalize long walks ($c_n \geq 0$ decreases with n)
  - Be a convergent series
  - Lead to a matrix function

New set of measures: matrix resolvent

- Use $(N - 1)^{n-1}$ as the weight, and $c_0 = N - 1$. After some math, we get

$$f(x) = (N-1)\left(1 - \frac{x}{N-1}\right)^{-1}$$

rescale to

$$f(x) = \left(1 - \frac{x}{N-1}\right)^{-1}$$

- Resolvent centrality

$$\sum_{k=1}^{N} \frac{N-1}{N-1-\lambda_k} \mathbf{x}_i^{[k]\,2}$$

- Resolvent communicability

$$\sum_{k=1}^{N} \frac{N-1}{N-1-\lambda_k} \mathbf{x}_i^{[k]} \mathbf{x}_j^{[k]}$$

- Resolvent betweenness

$$\frac{1}{(N-1)^2 - (N-1)} \sum \sum_{i \neq j, i \neq r, j \neq r} \frac{(f(A)_{ij} - f(A - E(r))_{ij})}{f(A)_{ij}}$$

Outline

- Motivation and background
- Measuring network with matrix exponential
- New set of measures: matrix resolvent
- Relation with graph Laplacian and spectral clustering
- Resolvent vs. exponential
- Experiment
- Conclusion

Relation with graph Laplacian and spectral clustering

- Partition the nodes into two groups, where nodes in one group share more edges whereas nodes across the two groups share less edges.
  - Define x_i = ½ if node *i* is in group A and x_i = -½ if node *i* is in group B
  - Solve for

$$\min_{\mathbf{x}\in\mathbb{R}^N, \|\mathbf{x}\|_2=1, \sum_{i=1}^N \mathbf{x}_i=0} \sum_{i=1}^{N}\sum_{j=1}^{N}(\mathbf{x}_i - \mathbf{x}_j)^2 a_{ij}$$

Number of edges across the two groups

  Set $||x||_2 = 1$ to eliminate trivial solution ($x = 0$); set $\sum_{i=1}^{N} x_i = 1$ to avoid built-in redundancy

  - Let $D = diag(deg_i)$, and rewrite

$$\min_{\mathbf{x}\in\mathbb{R}^N, \|\mathbf{x}\|_2=1, \sum_{i=1}^N \mathbf{x}_i=0} \mathbf{x}^T (D - A)\,\mathbf{x}.$$

  - *(D – A)* is the graph Laplacian. The solution is the eigenvector ($v^{[2]}$) with the second smallest eigenvalue (μ_2), which is referred as Fiedler vector.

Relation with graph Laplacian and spectral clustering

- For two nodes *i* and *j*, given the Fiedler vector, $v^{[2]}$,
  - $v_i^{[2]} v_j^{[2]} > 0$, i and j in the same group. Larger $v_i^{[2]} v_j^{[2]}$ means i and j are more communicable.
  - $v_i^{[2]} v_j^{[2]} < 0$, i and j in different groups. Smaller $v_i^{[2]} v_j^{[2]}$ means i and j are less communicable.
- In regular graph case with monotonic *f*, where the degree is uniform ($deg_i \equiv deg$), graph Laplacian becomes $deg\ I - A$ with eigenvalues $\boldsymbol{\mu}_i = deg - \lambda_i$, and eigenvectors $x^{[i]} = v^{[i]}$.
  - Dominant eigenvector ($k = 1$) contains no information as $x^{[1]} = e$.
  - Next dominate term is $v^{[2]} v^{[2]T}$, which is graph Laplacian clustering.
- In regular graph with monotonic *f*, $\min x^T (D - A)\, x = \min x^T (deg\ I - A)\, x \rightarrow \max x^T A\, x$.
  - Need x_i and x_j to be large and have the same sign → nodes that are more connected are farther from the origin.
  - $(v_i^{[2]})^2$ measures the well-connectedness of node *i* (same expression as *f*-centrality).

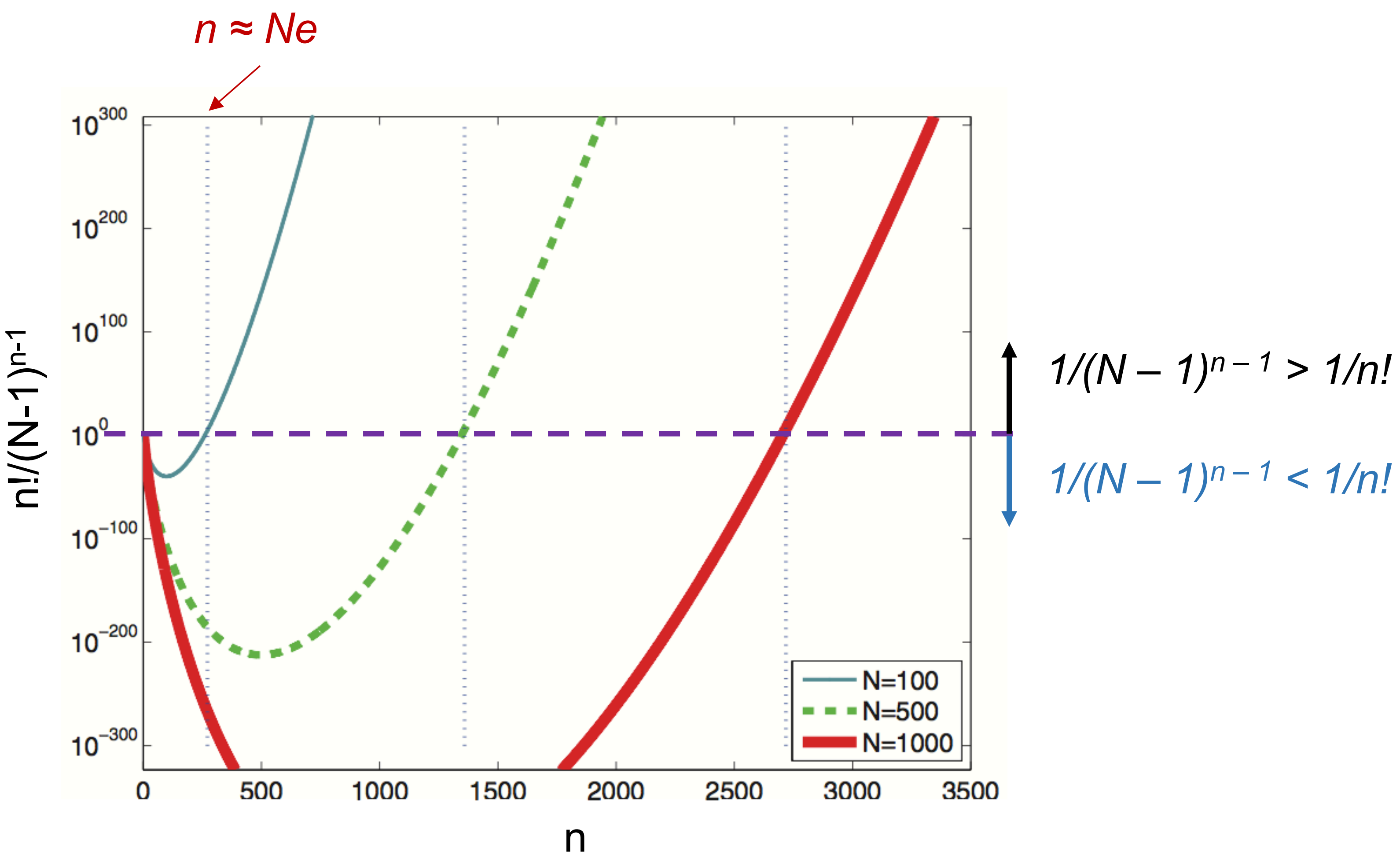
Outline

- Motivation and background
- Measuring network with matrix exponential
- New set of measures: matrix resolvent
- Relation with graph Laplacian and spectral clustering
- Resolvent vs. exponential
- Experiment
- Conclusion

Resolvent vs. exponential

- Exponential: $c_n = 1/n!$
- Resolvent: $c_n = 1/(N-1)^{n-1}$

- *n!* is comparable to $(N-1)^{n-1}$ when $n \approx Ne$.



- Exponential measures penalize less on short walks ($n < Ne$), whereas resolvent measures penalize less on long walks ($n > Ne$)

Resolvent vs. exponential

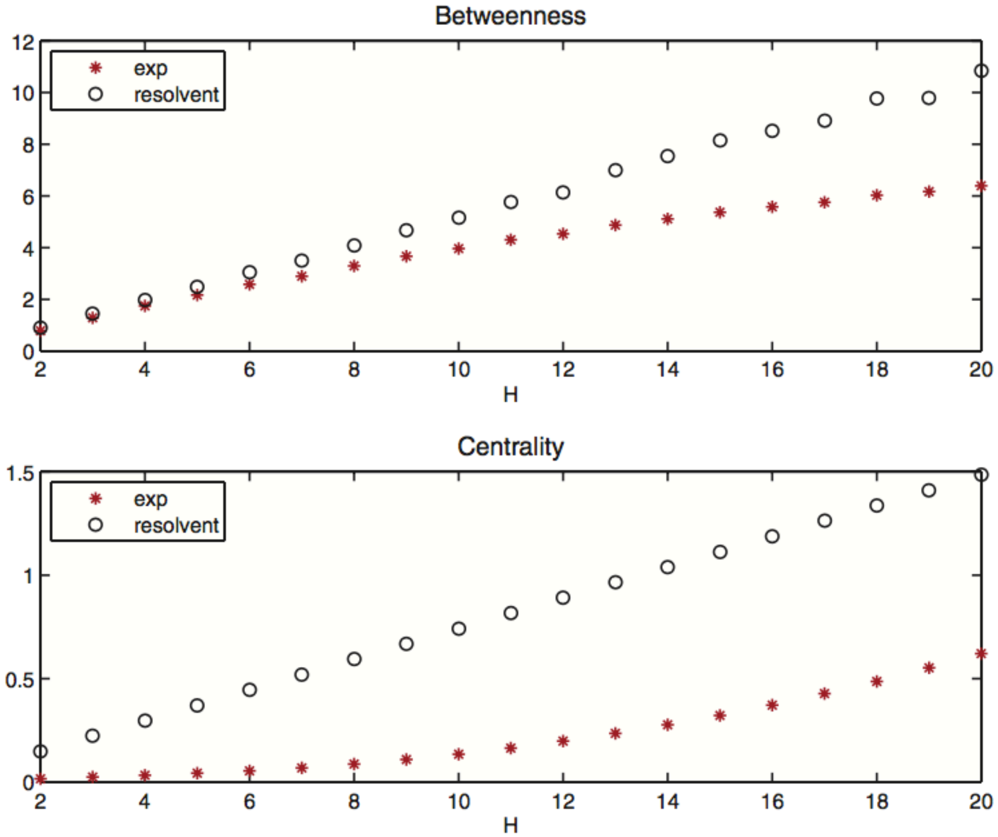
- The contribution from paths with length ≥ N to the resolvent communicability is bounded:

$$\sum_{n=N}^{\infty}\left(\frac{A^n}{(N-1)^{n-1}}\right)_{ij} \le \left\|\sum_{n=N}^{\infty}\frac{A^n}{(N-1)^{n-1}}\right\|_{\infty}$$
$$\le \sum_{n=N}^{\infty}\frac{\|A^n\|_{\infty}}{(N-1)^{n-1}}$$
$$\le \sum_{n=N}^{\infty}\frac{\| A \|_{\infty}^{n}}{(N-1)^{n-1}}$$
$$\le \frac{(N-1)^2}{N-1-\| A \|_{\infty}}\left(\frac{\| A \|_{\infty}}{N-1}\right)^{N}$$

- As $||A||_{\infty} << N$, we conclude that the contribution of walks with length $O(N)$ to the resolvent communicability is negligible.

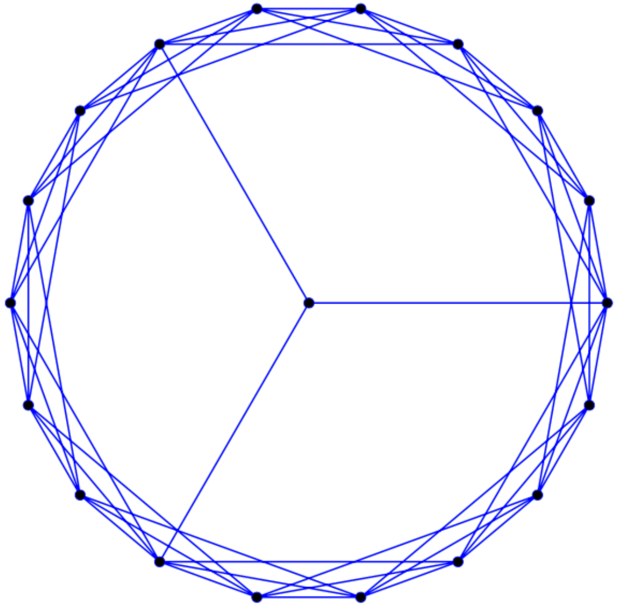
Ring of nodes with a hub

- Consider a periodic ring network. Each node is connected to *K* nodes clockwise and counterclockwise. Introduce a "hub" nodes that connects to *H* equally separated ring nodes.
- Consider such a network with *N = 200, K = 6*, and *H = 2 ~ 20*
- Hub removal destroys shortcuts → more effect on resolvent communicablility
- Resolvent centrality is more similar to node degree (ratio approaches to *1* as *H* approaches 13 ≈ *2K*)



Betweenness and centrality of hub node relative to a ring node

N = 18, K = 3, H = 3



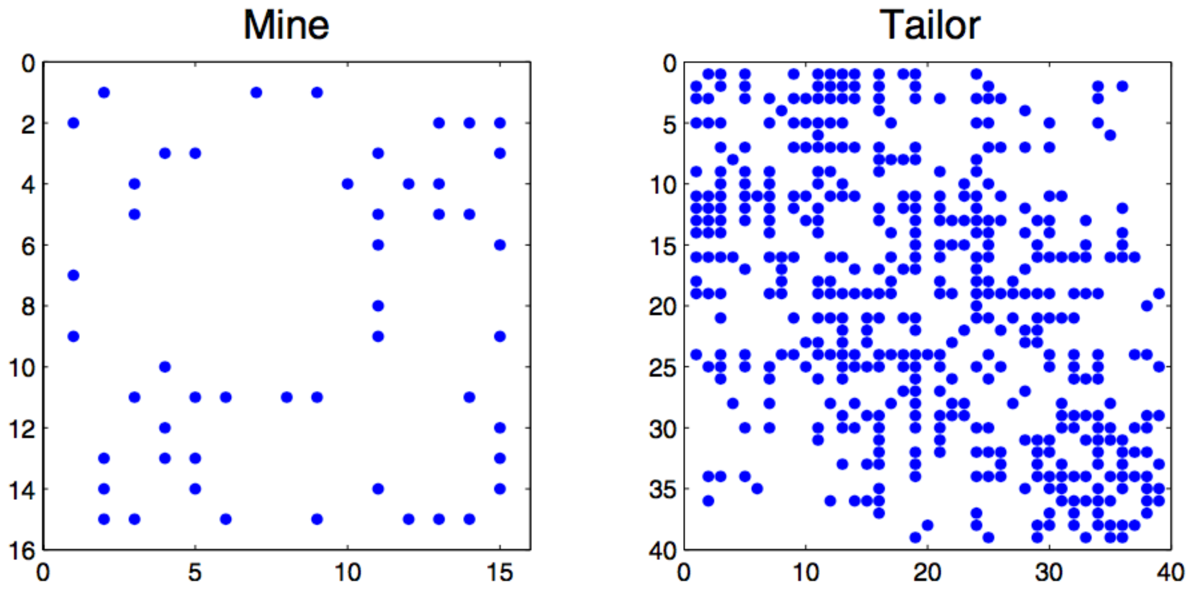
Outline

- Motivation and background
- Measuring network with matrix exponential
- New set of measures: matrix resolvent
- Relation with graph Laplacian and spectral clustering
- Resolvent vs. exponential
- Experiment
- Conclusion

Experiment with small scale data

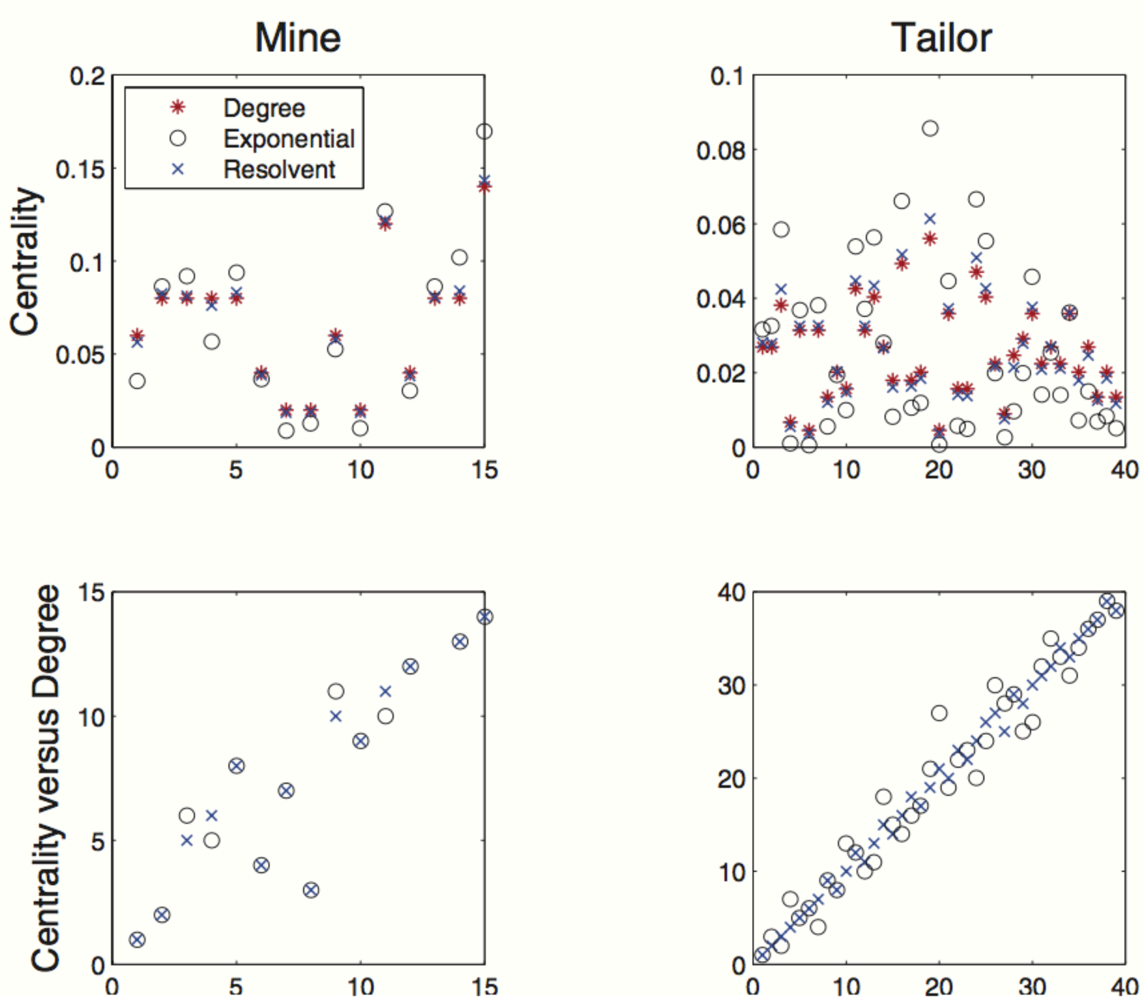
- Data set 1 – interactions (talk, joking, etc.) among15 mine workers in Zambia
- Data set 2 – assistant-level interactions among 40 individuals in a tailor's shop in Zambia.

Adjacency matrices



Experiment with small scale data

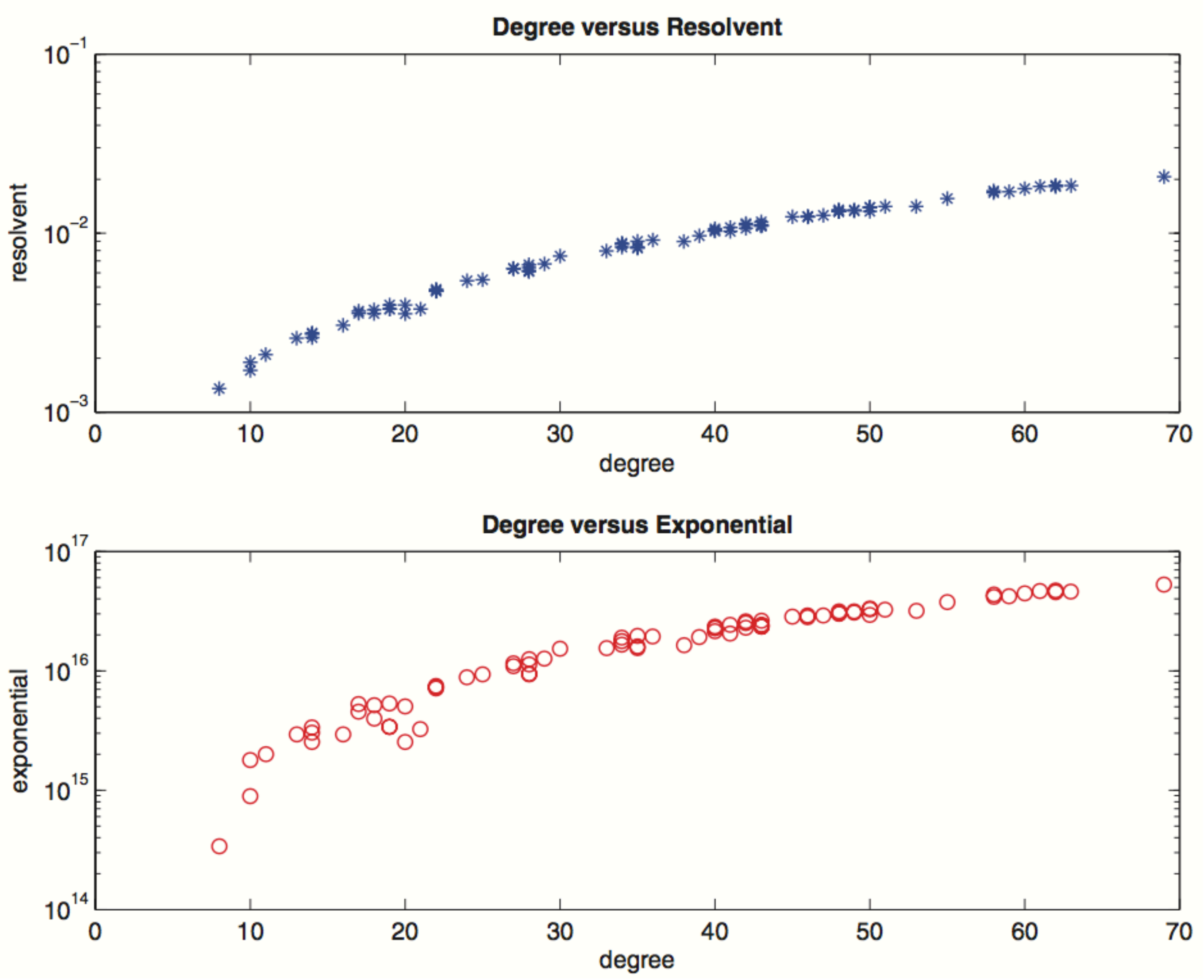
- Resolvent centrality is close to degree, and usually locates in the middle of degree and exponential centrality.
- The three measures differ more in the mine case (smaller, sparser).
- Requires arbitrary tie-breaking if nodes have the same degree.
- Resolvent is in the middle of exponential and degree, and is more close to degree.



Position (i, j) represents the node is ranked i in degree and j in exponential (circle) or resolvent (cross) centrality

Experiment with large scale data

- Food web: 81 nodes representing marine species.
- Resolvent measure is between the degree and exponential measures, and is more close to exponential.
- Ordering of the first 10 nodes in centrality ranking (high to low) is different in each measure.
- "≡" indicates a tie-breaking

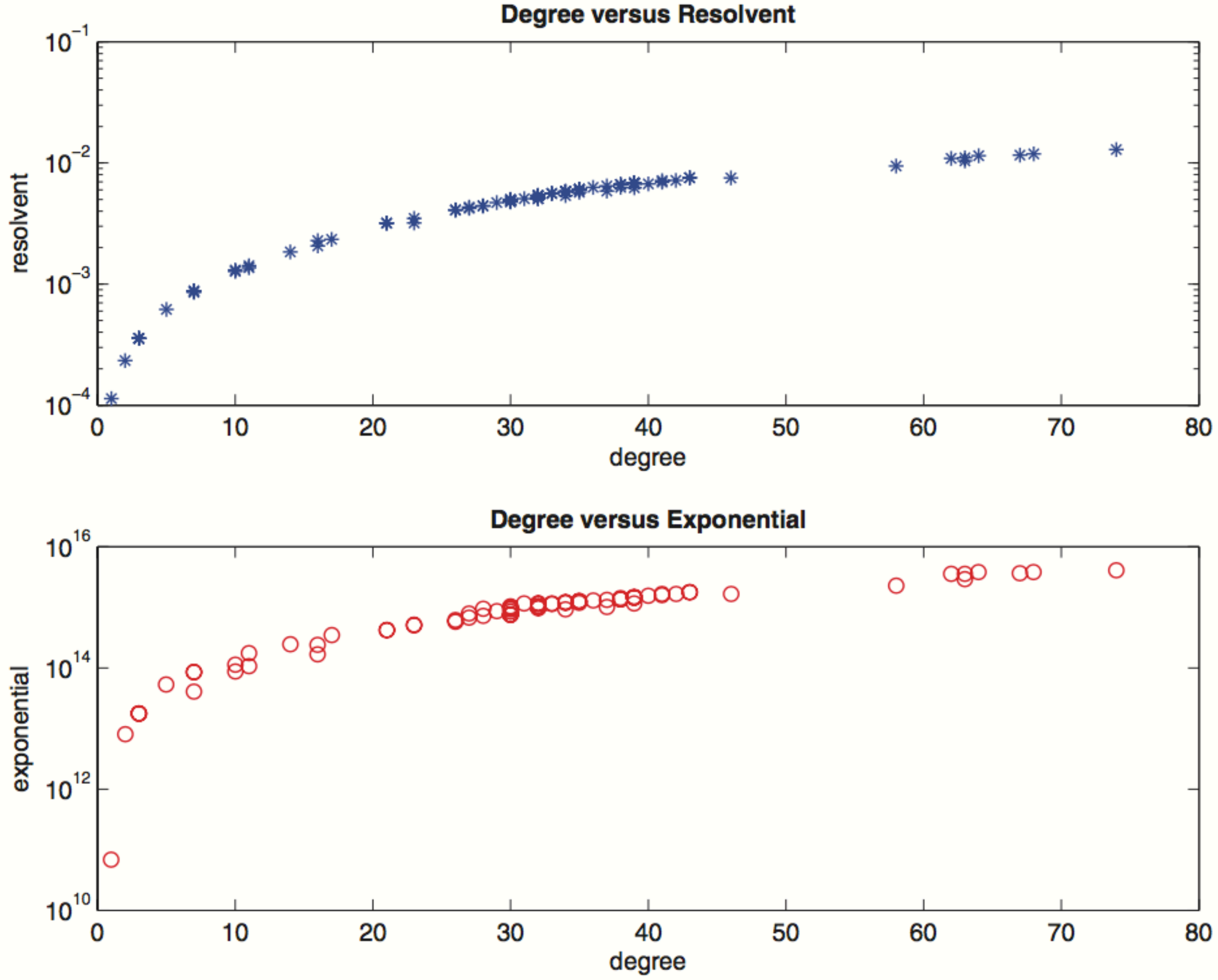


Increase centrality →

$$\begin{aligned} \textit{degree:}\quad & 50 < 49 \equiv 68 < 69 < 39 < 48 < 40 \equiv 41 < 42 < 43, \\ \textit{resolvent:}\quad & 50 < 49 < 69 < 68 < 39 < 41 < 48 < 42 < 40 < 43, \\ \textit{exponential:}\quad & 50 < 49 < 69 < 68 < 39 < 41 < 42 < 48 < 40 < 43. \end{aligned}$$

Experiment with large scale data

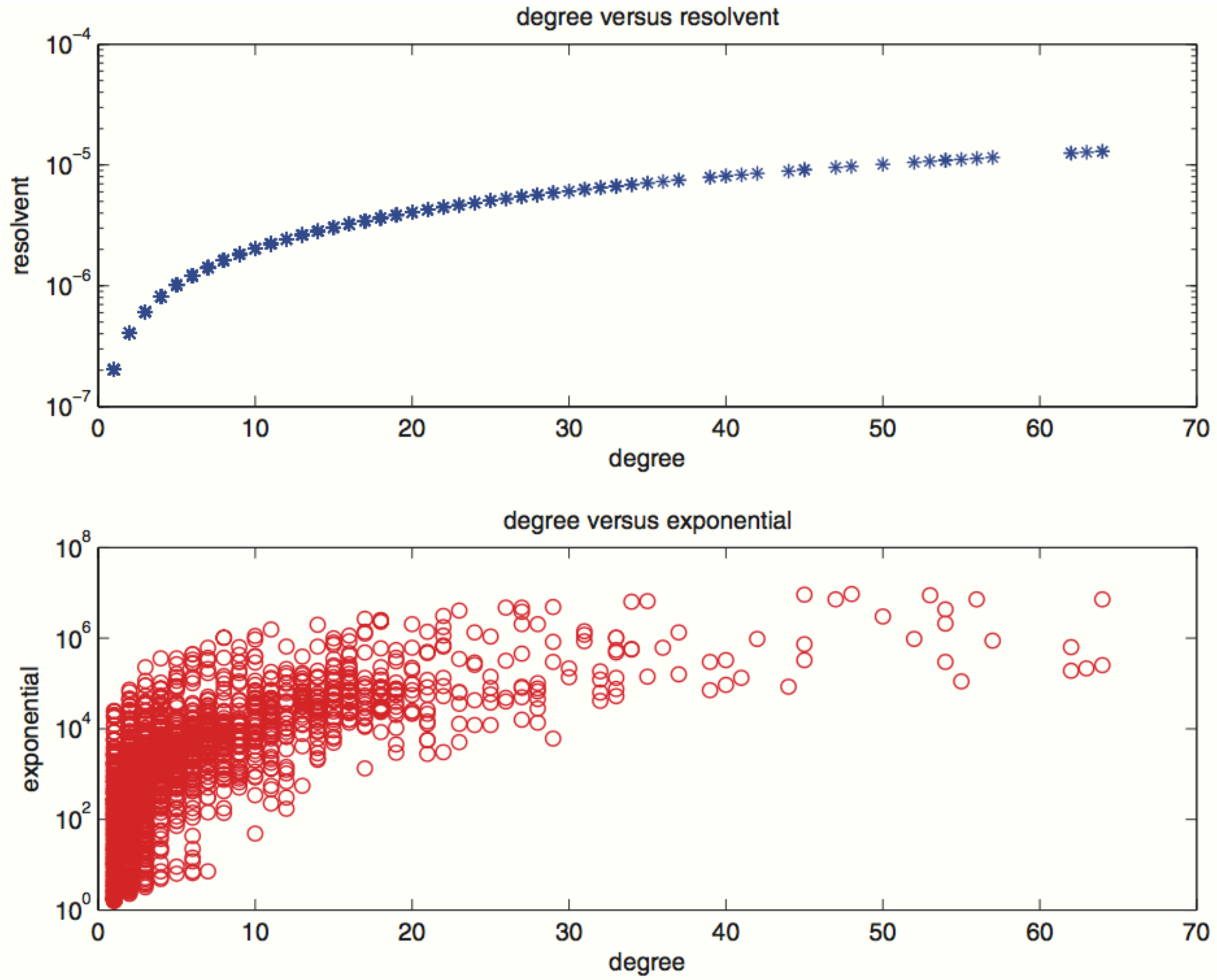
- Network of macaque cortical connectivity: 95 nodes representing regions in brain and edges are physical connections
- The three rankings are generally consistence. Difference appears at lower positions



$$
\begin{aligned}
\textit{degree:}\ & 15 < 94 < 65 < 58 < 31 \equiv 39 < 38 < 59 < 93 < 68,\\
\textit{resolvent:}\ & 13 < 6 < 65 < 39 < 58 < 31 < 38 < 59 < 93 < 68,\\
\textit{exponential:}\ & 11 < 12 < 65 < 39 < 58 < 31 < 59 < 38 < 93 < 68.
\end{aligned}
$$

Experiment with large scale data

- Protein-protein interaction network: 2224 proteins are nodes, each edge denotes an physical interaction.
- Note that nodes 111, 607 and 1896 have the same degree. This makes degree and resolvent equivalent.



$$
\begin{aligned}
\textit{degree:}\ & 607 \equiv 1896 < 489 < 473 < 138 < 200 \equiv 739 < 1338 < 292 \equiv 535, \\
\textit{resolvent:}\ & 111 < 607 < 489 < 473 < 138 < 739 < 200 < 1338 < 535 < 292, \\
\textit{exponential:}\ & 1170 < 122 < 129 < 156 < 117 < 473 < 292 < 242 < 126 < 427.
\end{aligned}
$$

Outline

- Motivation and background
- Measuring network with matrix exponential
- New set of measures: matrix resolvent
- Relation with graph Laplacian and spectral clustering
- Resolvent vs. exponential
- Experiment
- Conclusion

Conclusion

- Resolvent centrality is typically in the middle of degree and exponential centrality. Its value is often closer to degree in large networks.
- Resolvent measure has advantages of
  - Real-valued
  - Can yield analogous measures of communicability and betweenness
- Build the fundamentals of defining centrality, communicability, and betweenness using other matrix function