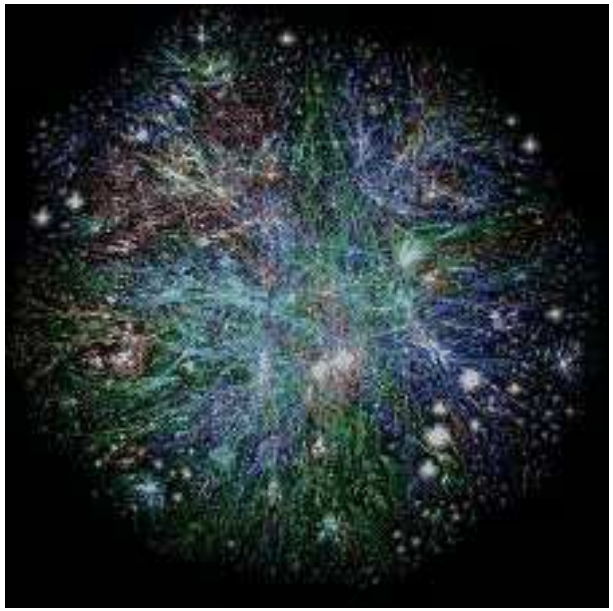
The Mathematics of the WEB

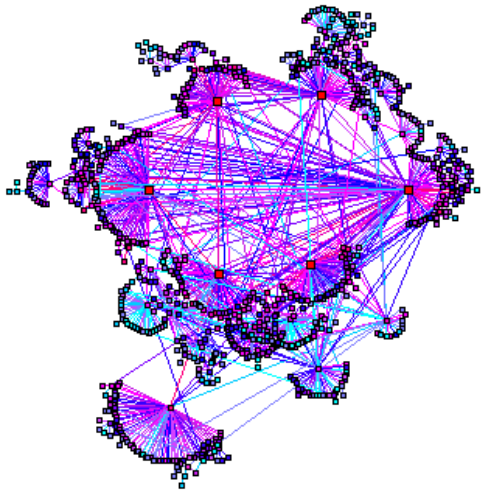
Edmundo de Souza e Silva[1]
Daniel Sadoc Menasche[2]

Federal University of Rio de Janeiro
[1]Systems Engineering and Computer Science Department, COPPE
[2]Computer Science Department, Math Institute

2011

Internet





- The Internet and the Web: full of intersting real problems
- Needs mathematical tools to understand/solve these

Outline

1. Introduction
2. Centrality
3. Clustering
4. Peer-to-Peer
5. Ending Remarks

Objective of the lecture

- To present a few recent practical problems of the WEB
- To show the mathematical foundations behind the solutions for the problems
- To motivate the students on the subject area
- The lecture is NOT:
  - To present solutions to the problems
  - To develop the mathematical tools

Objective of the lecture

- To present a few recent practical problems of the WEB
- To show the mathematical foundations behind the solutions for the problems
- To motivate the students on the subject area
- The lecture is NOT:
  - To present solutions to the problems
  - To develop the mathematical tools

Objective of the lecture

- To present a few recent practical problems of the WEB
- To show the mathematical foundations behind the solutions for the problems
- To motivate the students on the subject area
- The lecture is NOT:
  - To present solutions to the problems
  - To develop the mathematical tools

Objective of the lecture

- To present a few recent practical problems of the WEB
- To show the mathematical foundations behind the solutions for the problems
- To motivate the students on the subject area
- The lecture is NOT:
  - To present solutions to the problems
  - To develop the mathematical tools

Objective of the lecture

- To present a few recent practical problems of the WEB
- To show the mathematical foundations behind the solutions for the problems
- To motivate the students on the subject area
- The lecture is NOT:
  - To present solutions to the problems
  - To develop the mathematical tools

Last Lecture

- Introduced a few problems to motivate the importance of mathematical tools for obtaining interesting practical results
- Touch on topics such as:
  - Page Ranking
  - Markov Chains
  - Random Walk on Graphs

Last Lecture

- Introduced a few problems to motivate the importance of mathematical tools for obtaining interesting practical results
- Touch on topics such as:
  - Page Ranking
  - Markov Chains
  - Random Walk on Graphs

Last Lecture

- Introduced a few problems to motivate the importance of mathematical tools for obtaining interesting practical results
- Touch on topics such as:
  - Page Ranking
  - Markov Chains
  - Random Walk on Graphs

Last Lecture

- Introduced a few problems to motivate the importance of mathematical tools for obtaining interesting practical results
- Touch on topics such as:
  - Page Ranking
  - Markov Chains
  - Random Walk on Graphs

Lecture

- Will continue with:
  - Centrality
  - Clustering
  - P2P scalability
- Cover Application examples, not theory!

Lecture

- Will continue with:
  - Centrality
  - Clustering
  - P2P scalability
- Cover Application examples, not theory!

Outline

1. Introduction
2. Centrality
3. Clustering
4. Peer-to-Peer
5. Ending Remarks

Recall from last lecture

- Discussed the notion of PageRank
- PageRank is a metric of **centrality**
- Today's lecture: discuss other centrality metrics

Centrality

- Over the years researchers have introduced a large number of centrality indices
- These indices: measure of the importance of the vertices in a network
- Indices were valuable in the analysis and understanding of the roles played by actors in social networks, citation networks, computer networks, biological networks. $\cdots$

Centrality

- Over the years researchers have introduced a large number of centrality indices
- These indices: measure of the importance of the vertices in a network
- Indices were valuable in the analysis and understanding of the roles played by actors in social networks, citation networks, computer networks, biological networks. $\cdots$

Centrality

- Over the years researchers have introduced a large number of centrality indices
- These indices: measure of the importance of the vertices in a network
- Indices were valuable in the analysis and understanding of the roles played by actors in social networks, citation networks, computer networks, biological networks. $\cdots$

Applications

- Spreaders of disease in biological networks
- Key actors in terrorist networks
- Efficient marketing targets in social networks

Centrality

- A simple centrality measure: degree: is the number of edges incident on a vertex in a network.
  - In some sense it measures the popularity of an actor
- Another centrality measure: closeness: the mean shortest path distance between a node and all other nodes reachable from it.
  - Ex: measure of how long it takes from an information to spread from a node to all others in the network.

Centrality

- A simple centrality measure: degree: is the number of edges incident on a vertex in a network.
  - In some sense it measures the popularity of an actor
- Another centrality measure: closeness: the mean shortest path distance between a node and all other nodes reachable from it.
  - Ex: measure of how long it takes from an information to spread from a node to all others in the network.

Centrality

- A simple centrality measure: degree: is the number of edges incident on a vertex in a network.
  - In some sense it measures the popularity of an actor
- Another centrality measure: closeness: the mean shortest path distance between a node and all other nodes reachable from it.
  - Ex: measure of how long it takes from an information to spread from a node to all others in the network.

Centrality

- A simple centrality measure: degree: is the number of edges incident on a vertex in a network.
  - In some sense it measures the popularity of an actor
- Another centrality measure: closeness: the mean shortest path distance between a node and all other nodes reachable from it.
  - Ex: measure of how long it takes from an information to spread from a node to all others in the network.

Centrality
Betweeness

- Betweenness centrality: of a node *i* is defined as the fraction of shortest paths between pairs of vertices in a network that pass through *i*.
  - Measure of the extent to which an actor has control over information flowing between others.

Centrality
Betweeness

- Betweenness centrality: of a node *i* is defined as the fraction of shortest paths between pairs of vertices in a network that pass through *i*.
  - Measure of the extent to which an actor has control over information flowing between others.

Random walk and Centrality

- Centrality measures can be based on random walks (random walk samplying)
- *Power Centrality* can be seen in terms of random walks that have a fixed probability β of dying per step.
  The power centrality of vertex v_i is the expected number of times such a walker passes through v_i, averaged over all possible starting points for the walk.
- *Random Walk Betweenness* of a vertex v_i is the number of times that a random walker starting at s and ending at t passes through v_i along the way, averaged over all s and t.
- Linear algebra can be used for the solution of these measures
- Find the best technique to obtain accurate and efficient results

Random walk and Centrality

- Centrality measures can be based on random walks (random walk samplying)
- *Power Centrality* can be seen in terms of random walks that have a fixed probability β of dying per step.
  The power centrality of vertex v_i is the expected number of times such a walker passes through v_i, averaged over all possible starting points for the walk.
- *Random Walk Betweenness* of a vertex v_i is the number of times that a random walker starting at s and ending at t passes through v_i along the way, averaged over all s and t.
- Linear algebra can be used for the solution of these measures
- Find the best technique to obtain accurate and efficient results

Random walk and Centrality

- Centrality measures can be based on random walks (random walk samplying)
- *Power Centrality* can be seen in terms of random walks that have a fixed probability β of dying per step.
  The power centrality of vertex v_i is the expected number of times such a walker passes through v_i, averaged over all possible starting points for the walk.
- *Random Walk Betweenness* of a vertex v_i is the number of times that a random walker starting at s and ending at t passes through v_i along the way, averaged over all s and t.
- Linear algebra can be used for the solution of these measures
- Find the best technique to obtain accurate and efficient results

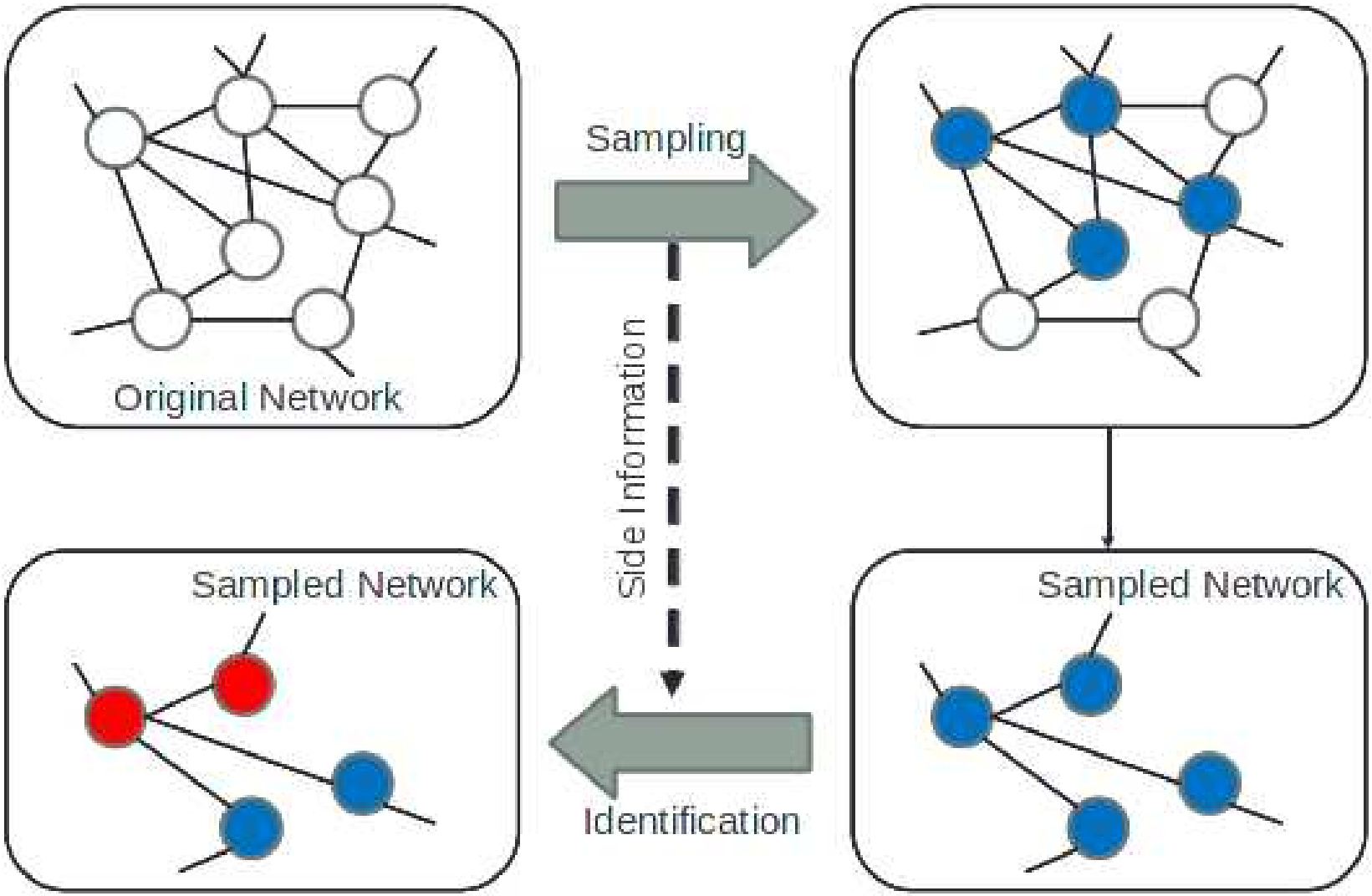
Random walk and Centrality

- Centrality measures can be based on random walks (random walk samplying)
- *Power Centrality* can be seen in terms of random walks that have a fixed probability β of dying per step.
  The power centrality of vertex v_i is the expected number of times such a walker passes through v_i, averaged over all possible starting points for the walk.
- *Random Walk Betweenness* of a vertex v_i is the number of times that a random walker starting at s and ending at t passes through v_i along the way, averaged over all s and t.
- Linear algebra can be used for the solution of these measures
- Find the best technique to obtain accurate and efficient results

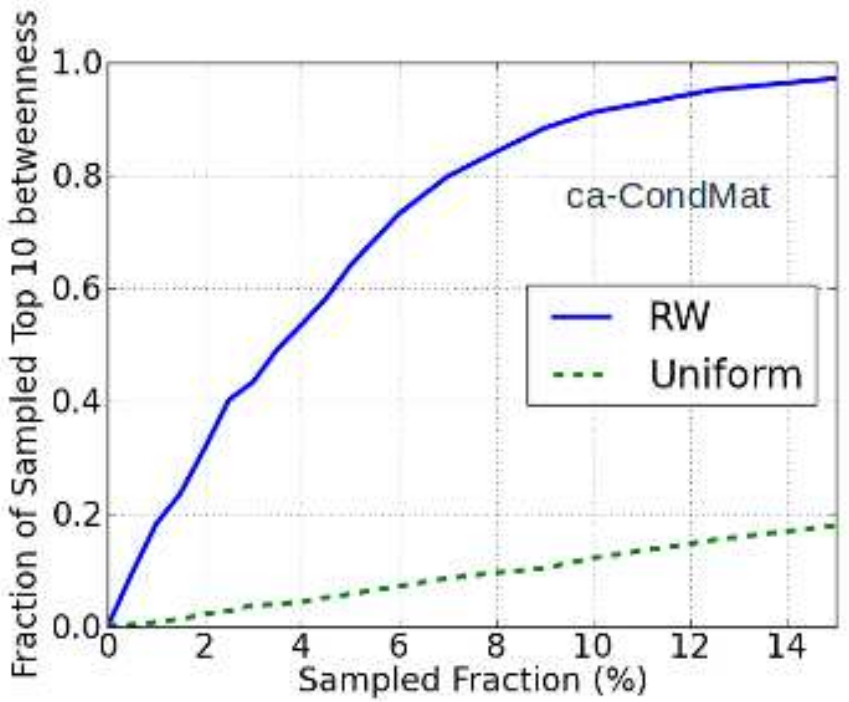
Random walk and Centrality

- Centrality measures can be based on random walks (random walk samplying)
- *Power Centrality* can be seen in terms of random walks that have a fixed probability β of dying per step.
  The power centrality of vertex v_i is the expected number of times such a walker passes through v_i, averaged over all possible starting points for the walk.
- *Random Walk Betweenness* of a vertex v_i is the number of times that a random walker starting at s and ending at t passes through v_i along the way, averaged over all s and t.
- Linear algebra can be used for the solution of these measures
- Find the best technique to obtain accurate and efficient results

Estimation Process

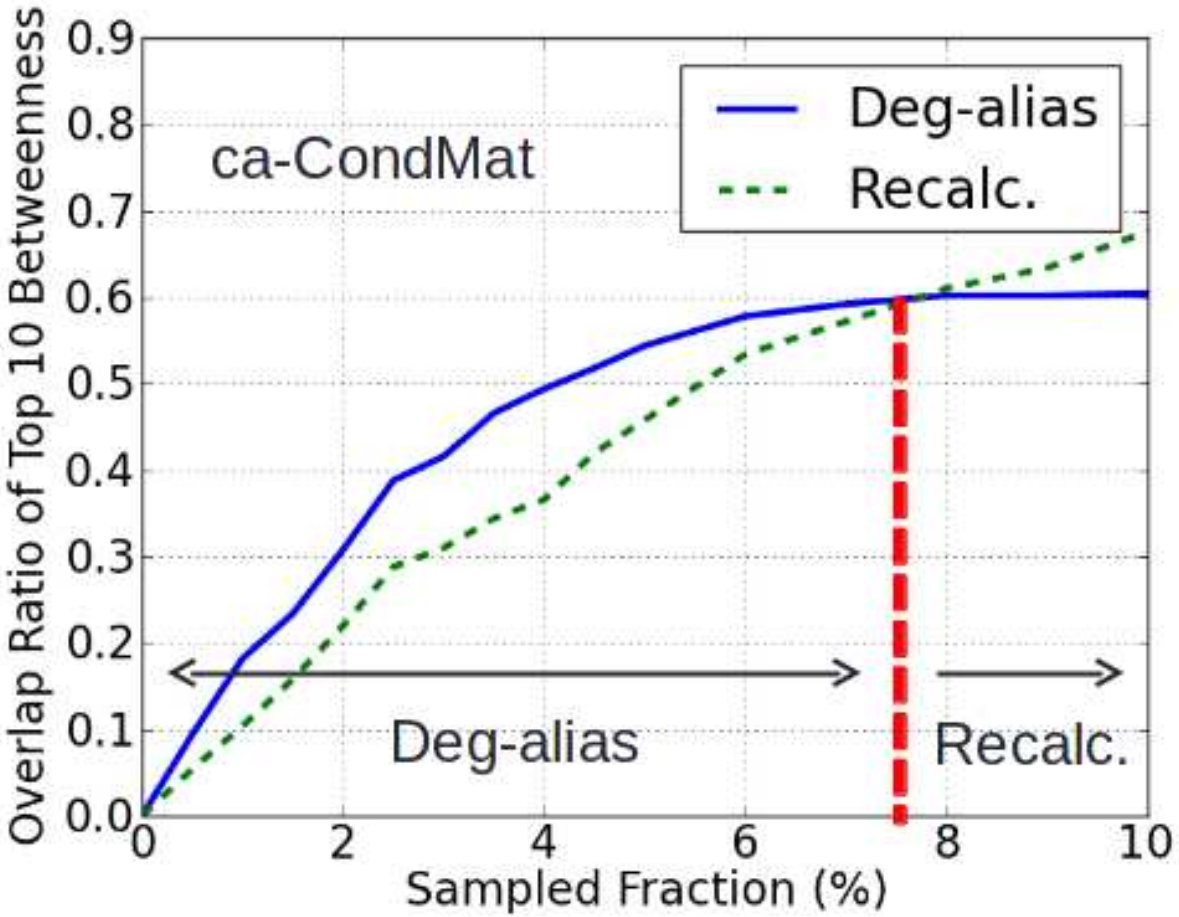


Sampling: Random Walk vs Uniform Sampling



Random walk sampling is more appropriate to collect most central nodes

Identification: Degree as Alias vs Recalculation



Outline

1. Introduction
2. Centrality
3. Clustering
4. Peer-to-Peer
5. Ending Remarks

Clustering
Basics

- Clustering is one of the most widely used techniques for exploratory data analysis
- One tries to identify groups of *similar behavior* in the studied dataset.
- Data can be represented in form of the similarity graph $G = (V, E)$. Each vertex represents a data point. Two vertices are connected if the similarity s_{ij} between the corresponding data points x_i and x_j is larger than a given threshold, and the edge is weighted by s_{ij}

Clustering
Basics

- Clustering is one of the most widely used techniques for exploratory data analysis
- One tries to identify groups of *similar behavior* in the studied dataset.
- Data can be represented in form of the similarity graph $G = (V, E)$. Each vertex represents a data point. Two vertices are connected if the similarity s_{ij} between the corresponding data points x_i and x_j is larger than a given threshold, and the edge is weighted by s_{ij}

Clustering
Basics

- Clustering is one of the most widely used techniques for exploratory data analysis
- One tries to identify groups of *similar behavior* in the studied dataset.
- Data can be represented in form of the similarity graph $G = (V, E)$. Each vertex represents a data point. Two vertices are connected if the similarity s_{ij} between the corresponding data points x_i and x_j is larger than a given threshold, and the edge is weighted by s_{ij}

The Problem

- The problem of clustering:
  - find a partition of the graph such that the edges between different groups have very low weights (points in different clusters are dissimilar from each other)
  - edges within a group have high weights (which means that points within the same cluster are similar to each other).
- Clustering can also be used to improve ranking algorithms:
  - create hierarchical clustering of results
  - this helps users to quickly find most appropriate category of results

The Problem

- The problem of clustering:
  - find a partition of the graph such that the edges between different groups have very low weights (points in different clusters are dissimilar from each other)
  - edges within a group have high weights (which means that points within the same cluster are similar to each other).
- Clustering can also be used to improve ranking algorithms:
  - create hierarchical clustering of results
  - this helps users to quickly find most appropriate category of results

The Problem

- The problem of clustering:
  - find a partition of the graph such that the edges between different groups have very low weights (points in different clusters are dissimilar from each other)
  - edges within a group have high weights (which means that points within the same cluster are similar to each other).
- Clustering can also be used to improve ranking algorithms:
  - create hierarchical clustering of results
  - this helps users to quickly find most appropriate category of results

The Problem

- The problem of clustering:
  - find a partition of the graph such that the edges between different groups have very low weights (points in different clusters are dissimilar from each other)
  - edges within a group have high weights (which means that points within the same cluster are similar to each other).
- Clustering can also be used to improve ranking algorithms:
  - create hierarchical clustering of results
  - this helps users to quickly find most appropriate category of results

Solutions

- The simplest and most direct way to construct a partition of the graph is to solve the mincut problem.
- relatively easy problem. But...
- In practice often does not lead to satisfactory partitions

Solutions

- The simplest and most direct way to construct a partition of the graph is to solve the mincut problem.
- relatively easy problem. But...
- In practice often does not lead to satisfactory partitions

Solutions

- The simplest and most direct way to construct a partition of the graph is to solve the mincut problem.
- relatively easy problem. But...
- In practice often does not lead to satisfactory partitions

Random walks

- Problem: find a partition of the graph such that the random walk stays long within the same cluster and seldom jumps between clusters.
- The transition probability **P** of jumping from vertex v_i to v_j is proportional to the edge weight w_{ij}: $p_{ij} := w_{ij}/d_i$.
- Eigenvectors and eigenvalues of **P** can be used to describe cluster properties of the graph.

Random walks

- Problem: find a partition of the graph such that the random walk stays long within the same cluster and seldom jumps between clusters.
- The transition probability **P** of jumping from vertex v_i to v_j is proportional to the edge weight w_{ij}: $p_{ij} := w_{ij}/d_i$.
- Eigenvectors and eigenvalues of **P** can be used to describe cluster properties of the graph.

Random walks

- Problem: find a partition of the graph such that the random walk stays long within the same cluster and seldom jumps between clusters.
- The transition probability **P** of jumping from vertex v_i to v_j is proportional to the edge weight w_{ij}: $p_{ij} := w_{ij}/d_i$.
- Eigenvectors and eigenvalues of **P** can be used to describe cluster properties of the graph.

Issues

- In practice one has to compute the first k eigenvectors of a potentially large graph matrix.
- But one can take advantage of sparseness and other propreties. (Linear algebra!)
- Choosing the number of clusters is also an issue for clustering algorithms

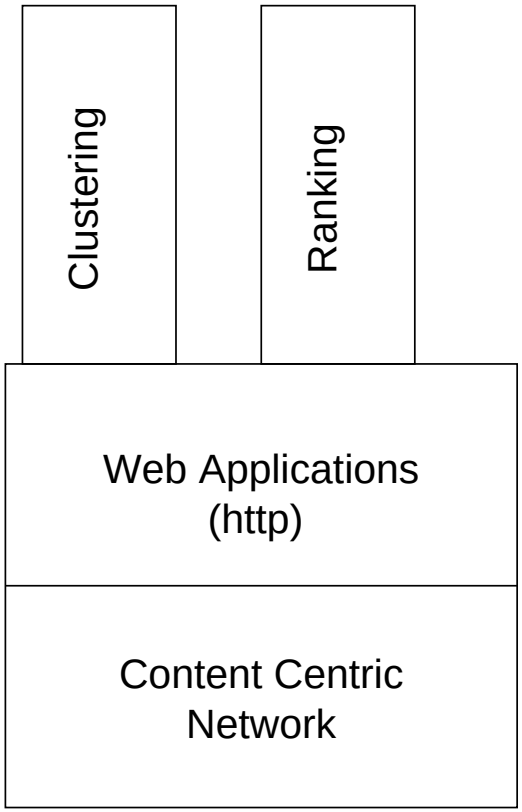
Issues

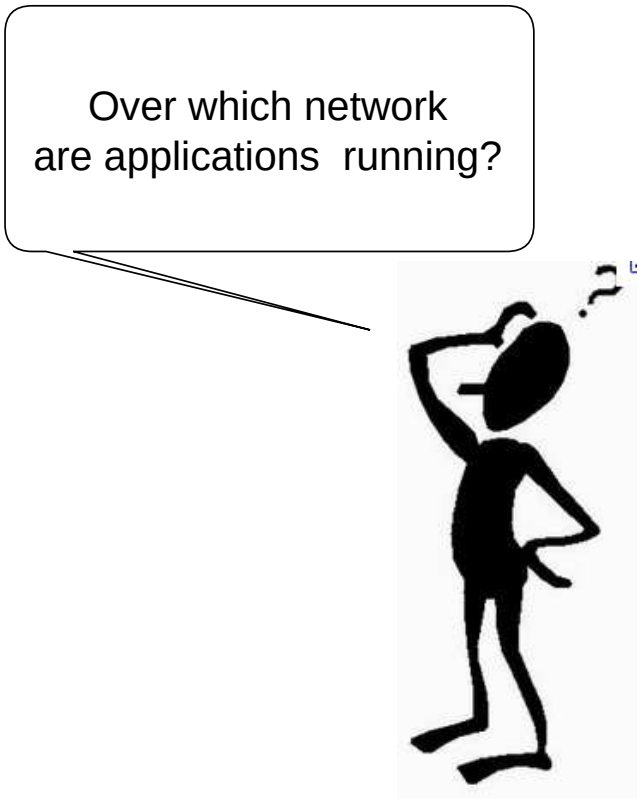
- In practice one has to compute the first k eigenvectors of a potentially large graph matrix.
- But one can take advantage of sparseness and other propreties. (Linear algebra!)
- Choosing the number of clusters is also an issue for clustering algorithms

Issues

- In practice one has to compute the first k eigenvectors of a potentially large graph matrix.
- But one can take advantage of sparseness and other propreties. (Linear algebra!)
- Choosing the number of clusters is also an issue for clustering algorithms

Applications and Network





Outline

1 Introduction

2 Centrality

3 Clustering

4 Peer-to-Peer

5 Ending Remarks

P2P Content Dissemination

- Why Smarter Information Sharing?
- Today's society: immense amount of information
  - Youtube: 34 hours of video uploaded every minute
  - 2008 Olympics: NBC Olympics served 75.5 million videos
  - World of Warcraft: 11.5 million subscribers worldwide
  - Wikipedia: 100K requests/s
  - Linux distribution: 100 GB/hour

P2P Content Dissemination

- Why Smarter Information Sharing?
- Today's society: immense amount of information
  - Youtube: 34 hours of video uploaded every minute
  - 2008 Olympics: NBC Olympics served 75.5 million videos
  - World of Warcraft: 11.5 million subscribers worldwide
  - Wikipedia: 100K requests/s
  - Linux distribution: 100 GB/hour

P2P Content Dissemination

- Goals: Distribute content accounting for
  - availability
  - scalability
  - robustness
- Solution
  - peer-to-peer systems

P2P Content Dissemination

- Goals: Distribute content accounting for
  - availability
  - scalability
  - robustness
- Solution
  - peer-to-peer systems

P2P Content Dissemination

- Mathematics of Peer to Peer Systems: Models, measurements and tools to leverage peer-to-peer swarming
- Account for:
  - searching
  - availability of content
  - scalability
  - robustness

Searching

- searching and construction of unstructured peer-to-peer (P2P) are yet another application example of random walks
- Random Walks for searching may achieve good results. It is an excellent candidate to simulate sampling for P2P networks
- It has also shown to be good for constructing and maintaining a P2P topology

Searching

- searching and construction of unstructured peer-to-peer (P2P) are yet another application example of random walks
- Random Walks for searching may achieve good results. It is an excellent candidate to simulate sampling for P2P networks
- It has also shown to be good for constructing and maintaining a P2P topology

Searching

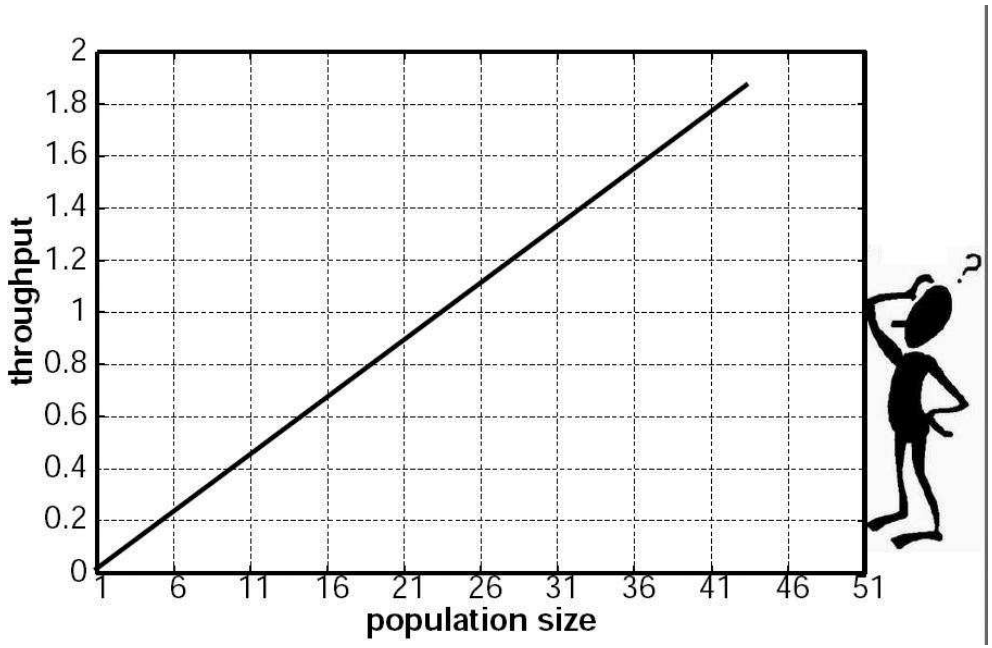
- searching and construction of unstructured peer-to-peer (P2P) are yet another application example of random walks
- Random Walks for searching may achieve good results. It is an excellent candidate to simulate sampling for P2P networks
- It has also shown to be good for constructing and maintaining a P2P topology

Other important Questions

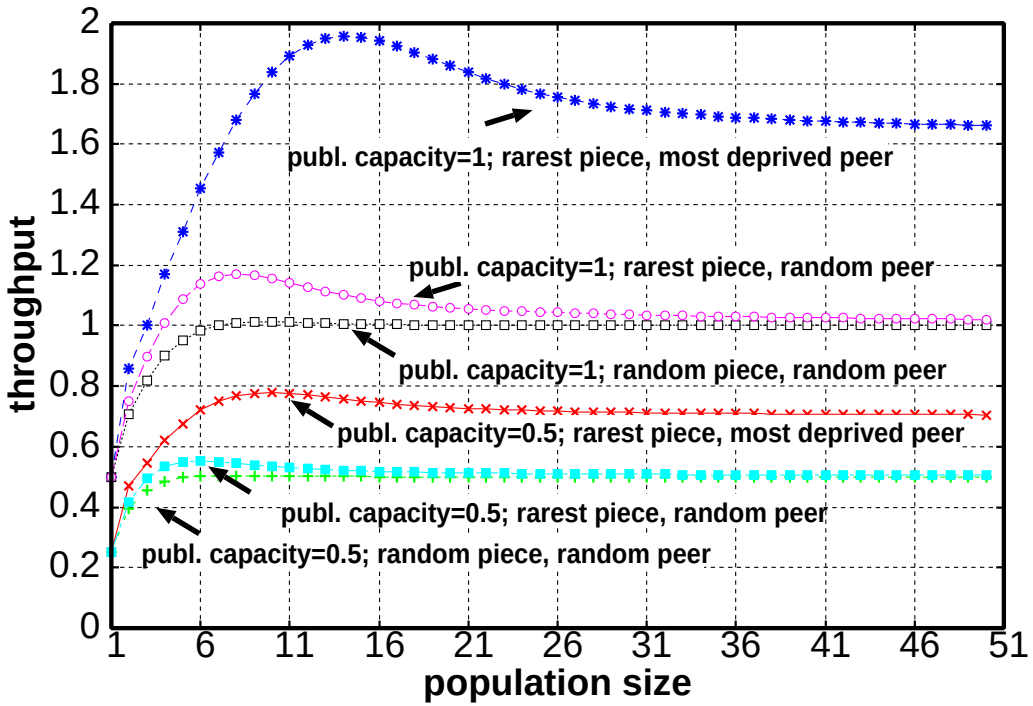
- **content availability**: BitTorrent great to disseminate popular content
  - what about niche content?
- **system robustness**: peers complete download even in abscence of publisher
  - what fraction of time is content available among peers?
- **reciprocity**: BitTorrent implements tit-for-tat; eMule implements credit based system
  - what are the advantages and disadvantages of each?
- **scalability**: system capacity scales with population size
  - what are fundamental limitations on scalability?

Scalability

- key property p2p swarming: every peer is server and client
- # of servers = # of peers
- therefore, service capacity (throughput) increases linearly with population size?



P2P Availability



Further Details

- Implications of Peer Selection Strategies by Publishers on the Performance And Stability of P2P Swarming Systems; Menasche, Rocha, Souza e Silva, Towsley, Leao
- Attend talk by Antonio Rocha, Wednesday 3.00 PM

Further Details

- Implications of Peer Selection Strategies by Publishers on the Performance And Stability of P2P Swarming Systems; Menasche, Rocha, Souza e Silva, Towsley, Leao
- Attend talk by Antonio Rocha, Wednesday 3.00 PM

Ending Remarks

- Lots of applications: we cover just a few to show that the importance of mathematical tools for solving interesting problems.
- One such useful tool is Random Walks $\rightarrow$ Markov chain analysis $\rightarrow$ linear algebra and probability theory.
- besides graph theory, of course!
- Without the theory one cannot solve a lot of real world problems.

Ending Remarks

- Lots of applications: we cover just a few to show that the importance of mathematical tools for solving interesting problems.
- One such useful tool is Random Walks $\rightarrow$ Markov chain analysis $\rightarrow$ linear algebra and probability theory.
- besides graph theory, of course!
- Without the theory one cannot solve a lot of real world problems.

Ending Remarks

- Lots of applications: we cover just a few to show that the importance of mathematical tools for solving interesting problems.
- One such useful tool is Random Walks $\rightarrow$ Markov chain analysis $\rightarrow$ linear algebra and probability theory.
- besides graph theory, of course!
- Without the theory one cannot solve a lot of real world problems.

Ending Remarks

- Lots of applications: we cover just a few to show that the importance of mathematical tools for solving interesting problems.
- One such useful tool is Random Walks $\rightarrow$ Markov chain analysis $\rightarrow$ linear algebra and probability theory.
- besides graph theory, of course!
- Without the theory one cannot solve a lot of real world problems.

Ending Remarks

We invite you to join a future class to learn

The Mathematics of the Web

Ending Remarks

We invite you to join a future class to learn

The Mathematics of the Web

THANKS