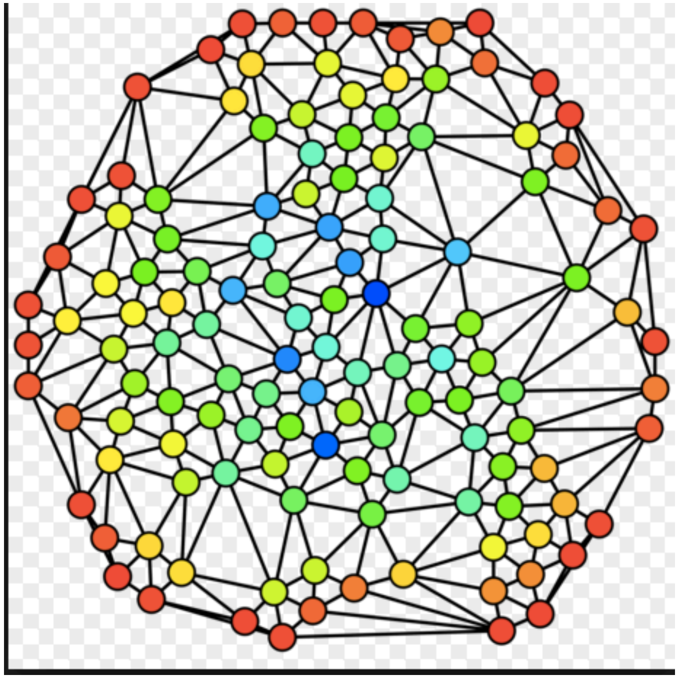
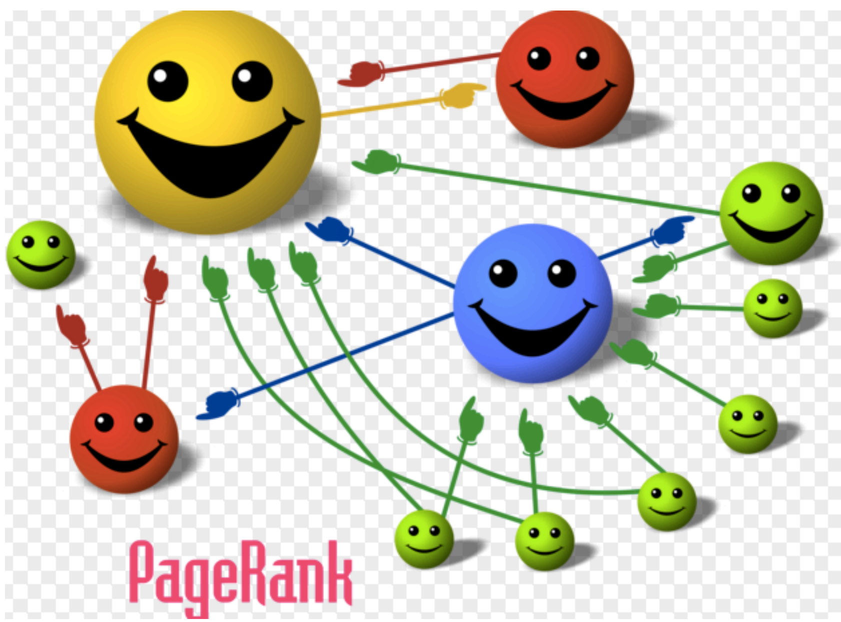


Complex Networks: Network Centrality

Ana Paula Couto

Computer Science Department

Universidade Federal de Minas Gerais



How we can identify the most important vertices within a network?

- The most influential person(s) in a social network
- Key infrastructure nodes in the Internet or urban networks
- Super-spreaders of disease

How we can identify the most important vertices within a network?

- Using the network topology
  - Degree
  - Betweenness
  - Closeness
  - Page Rank

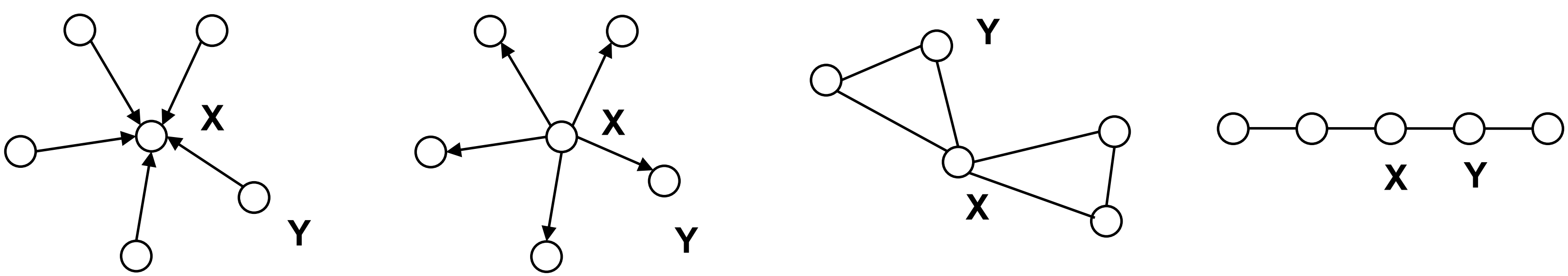
How we can identify the most important vertices within a network?

- Local Metrics
  - Based on the node's neighborhood
  - Low computational cost
    - Degree

- Global Metrics
  - All graph structure
  - High computational cost for large networks
    - Closeness, Pagerank

Different concepts and networks, different centrality levels

- X is more central than Y, for different networks and metrics



indegree outdegree betweenness closeness

What is the best centrality metric?

- How to evaluate the rank quality?
  - Ground-truth: context dependent
  - What is the goal of ranking?
- There is not the best metric!

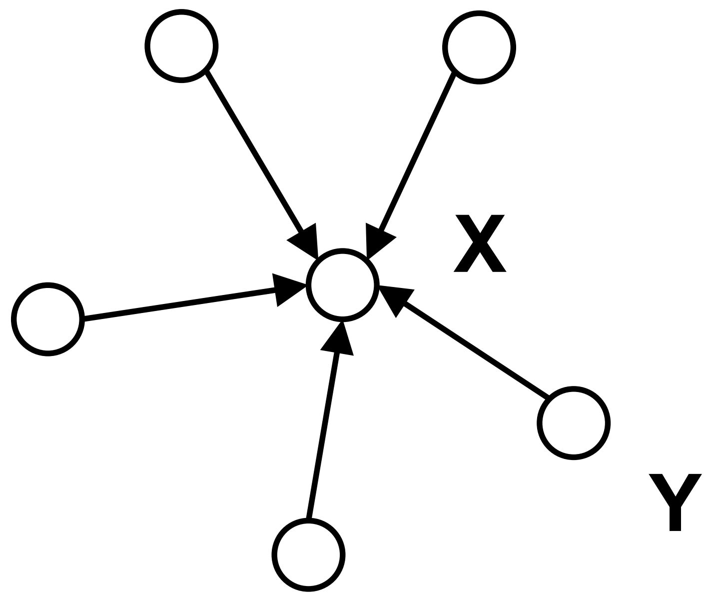
Degree centrality

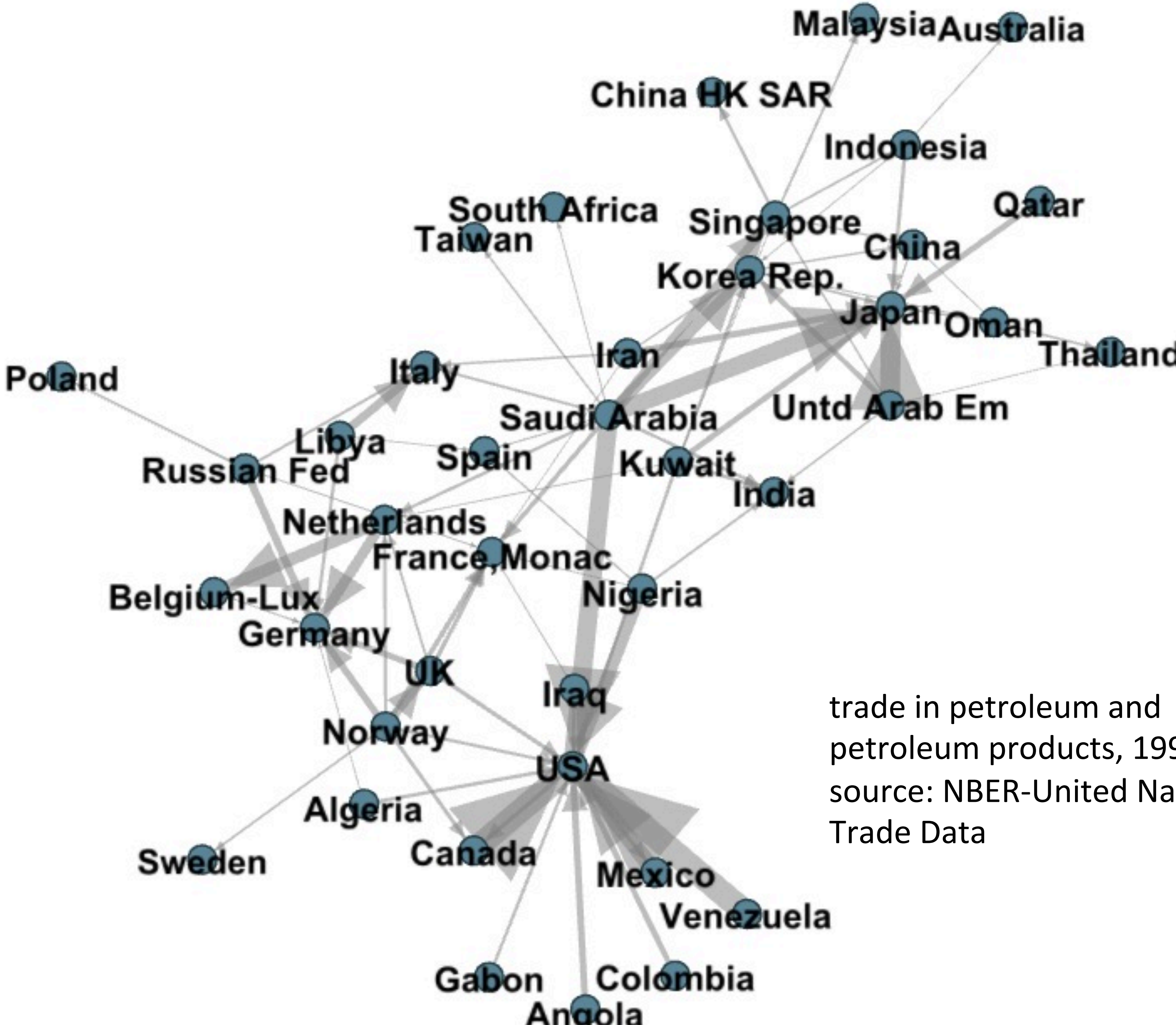
- Is the normalized node degree value

$$DC_i = k_i /(N-1)$$

- Directed graph
  - In-degree centrality
  - Out-degree centrality

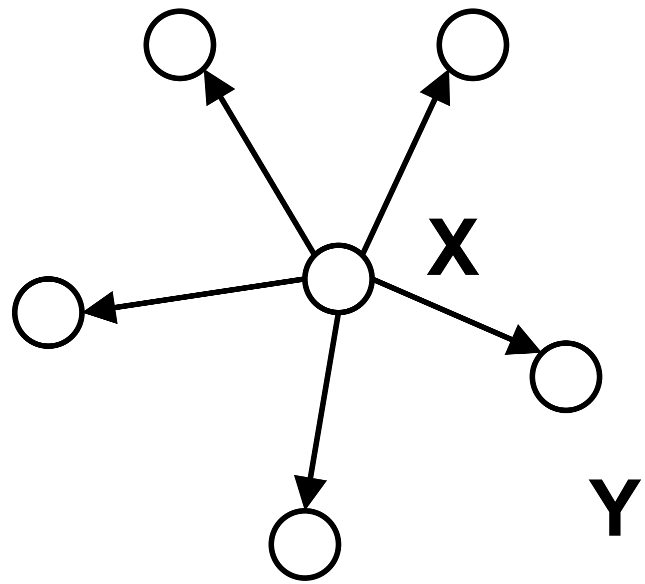
In-degree centrality

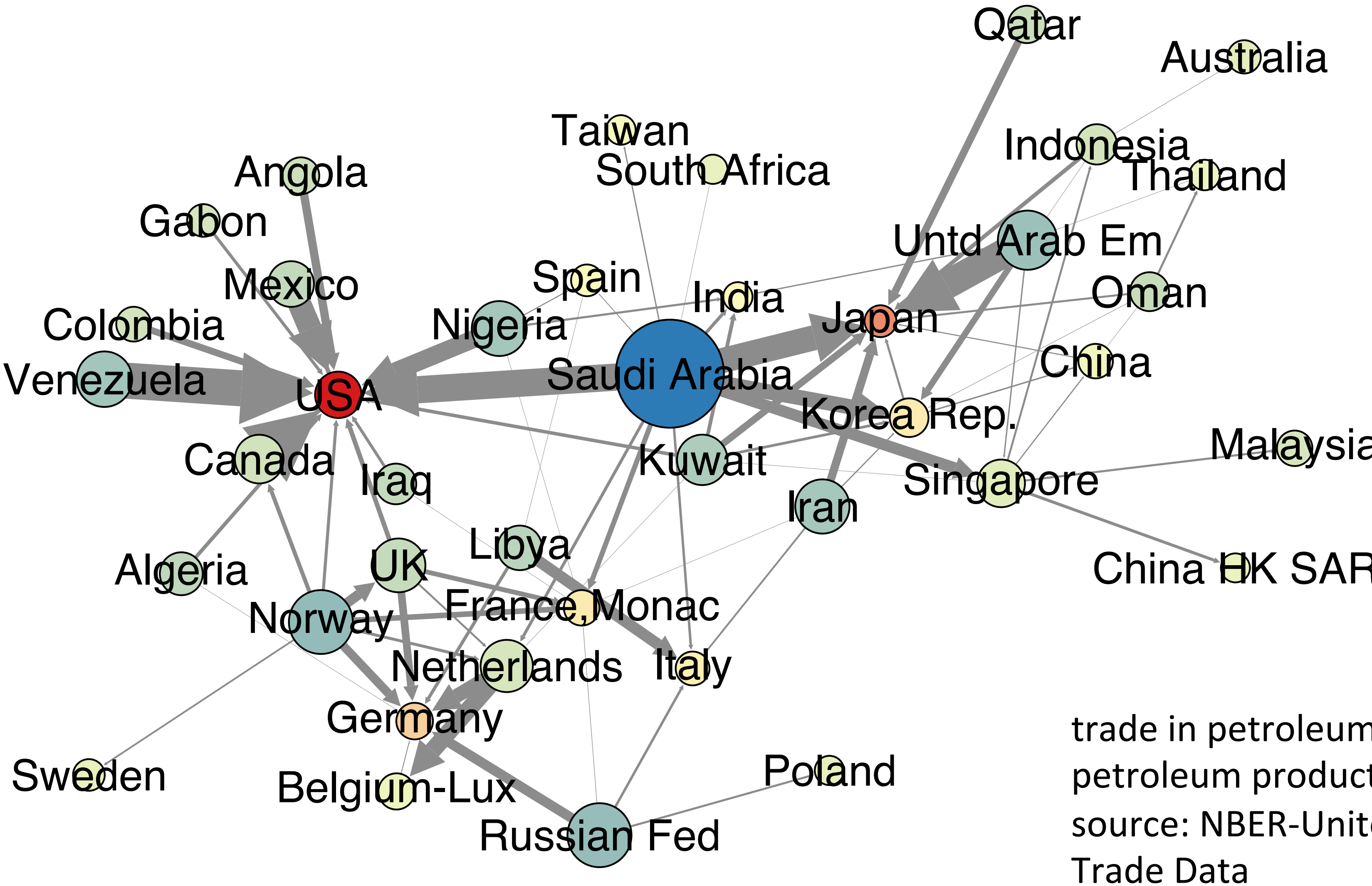




trade in petroleum and petroleum products, 1998, source: NBER-United Nations Trade Data

Out-degree centrality

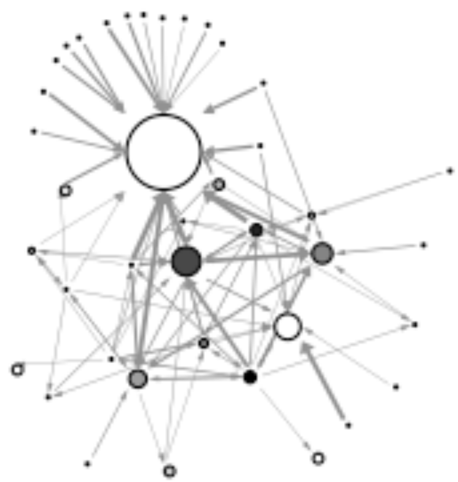




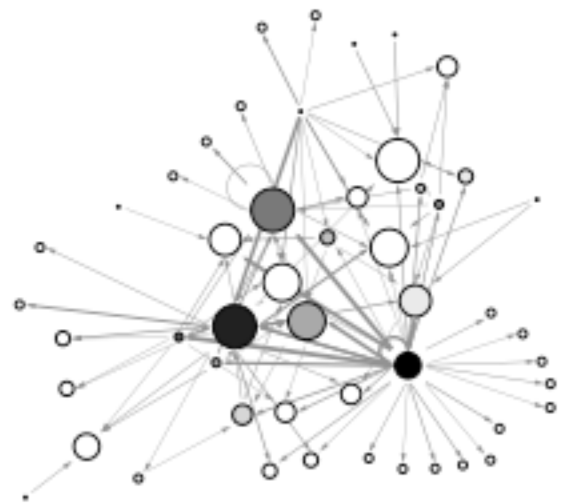
trade in petroleum and
petroleum products, 1998
source: NBER-United Natio
Trade Data

Real Network Example

example financial trading networks

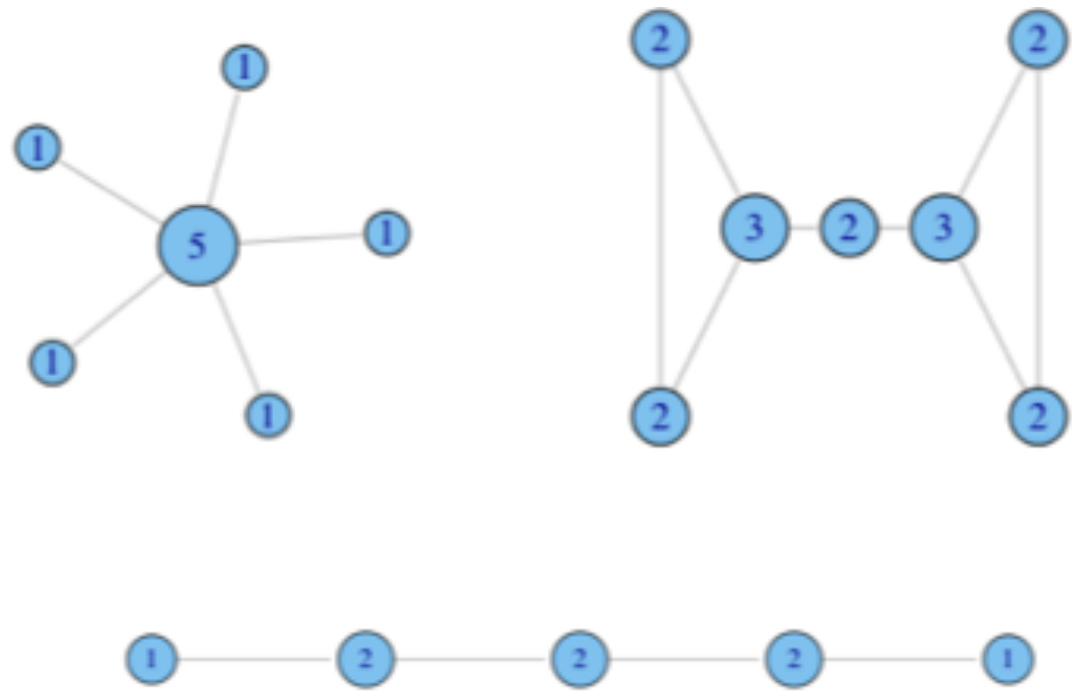


high in-centralization: one node buying from many others



low in-centralization: buying is more evenly distributed

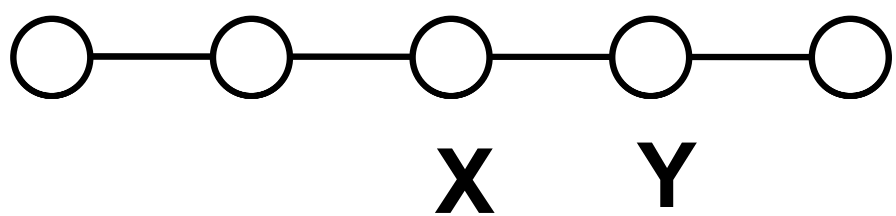
What node degree did not capture



- ability to broker between groups
- likelihood that information originating anywhere in the network reaches you

Betweenness

- **Intuition**: how many pairs of individuals would have to go through you in order to reach one another in the minimum number of hops?



Betweenness calculation

$$C_B(i) = \sum_{j<k} g_{jk}(i) / g_{jk}$$

Where g_{jk} = the number of geodesics connecting *jk*, and $g_{jk}(i)$ = the number of paths that actor *i* is on

Usually normalized by:

$$C'_B(i) = C_B(i) / \boxed{[(n-1)(n-2)/2]}$$

number of pairs of vertices excluding the vertex itself

Betweenness on toy networks

- non-normalized version:

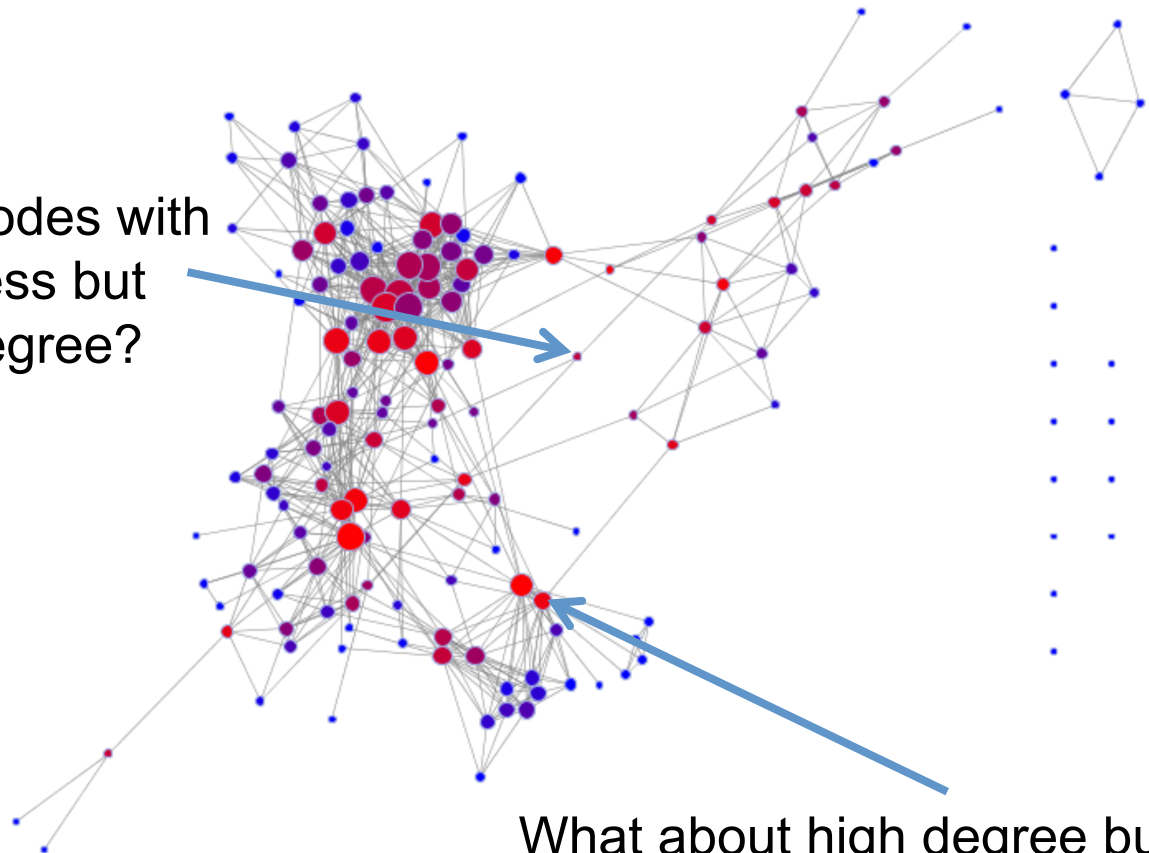
- A lies between no two other vertices
- B lies between A and 3 other vertices: C, D, and E
- C lies between 4 pairs of vertices (A,D),(A,E),(B,D),(B,E)

- note that there are no alternate paths for these pairs to take, so C gets full credit

Betweenness: Real example

Lada's old Facebook network: nodes are sized by degree, and colored by betweenness.

Can you spot nodes with high betweenness but relatively low degree?



What about high degree but relatively low betweenness?

Closeness

- What if it's not so important to have many direct friends?
- Or be "between" others
- But one still wants to be in the "middle" of things, not too far from the center

Closeness calculation

Closeness is based on the length of the average shortest path between a vertex and all vertices in the graph

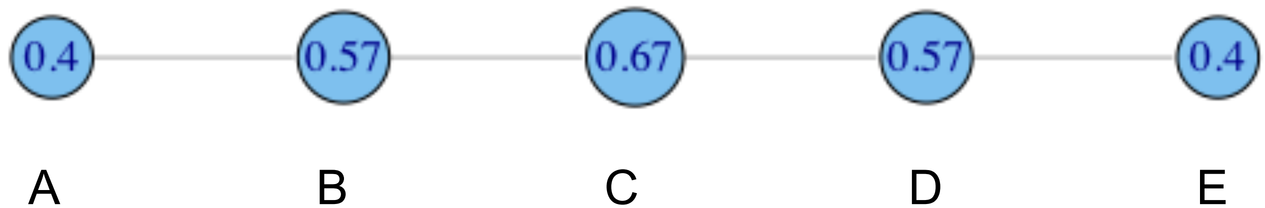
Closeness Centrality:

$$C_c(i) = \left[\sum_{j=1}^{N} d(i,j) \right]^{-1}$$

Normalized Closeness Centrality

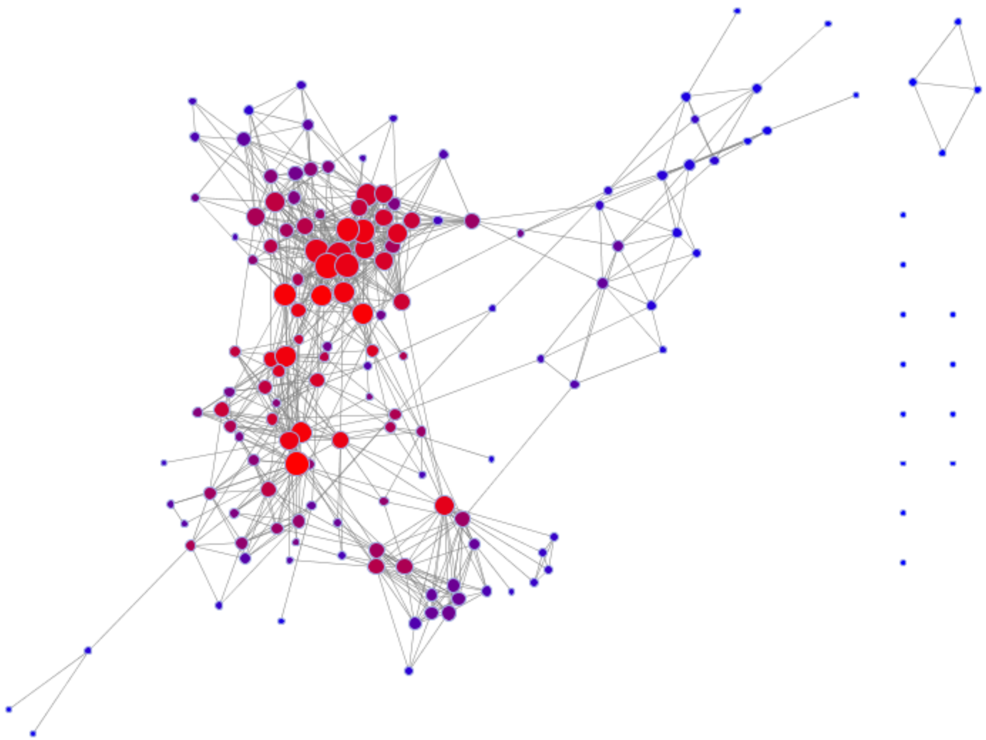
$$C_C^{'}(i) = (C_C(i))/(N-1)$$

Closeness on toy networks



$$C_c^{'}(A) = \left[\frac{\sum_{j=1}^{N} d(A,j)}{N-1}\right]^{-1} = \left[\frac{1+2+3+4}{4}\right]^{-1} = \left[\frac{10}{4}\right]^{-1} = 0.4$$

Closeness: Real Example



- **degree** (number of connections) denoted by size
- **closeness** (length of shortest path to all others) denoted by color

Prestige in directed social networks

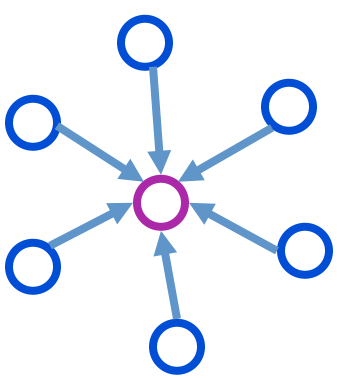
- when 'prestige' may be the right word
  - Admiration, influence, trust

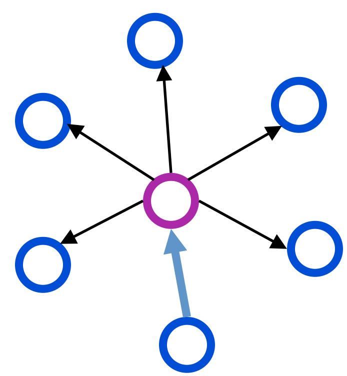
- directionality especially important in instances where ties may not be reciprocated (e.g. dining partners choice network)

- when 'prestige' may not be the right word
  - gives advice to (can reverse direction)
  - gives orders to (- " -)
  - lends money to (- " -)
  - dislikes
  - distrusts

Extensions of undirected degree centrality - prestige

- Degree centrality
  - In-degree centrality
    - a paper that is cited by many others has high prestige
    - a person nominated by many others for a reward has high prestige

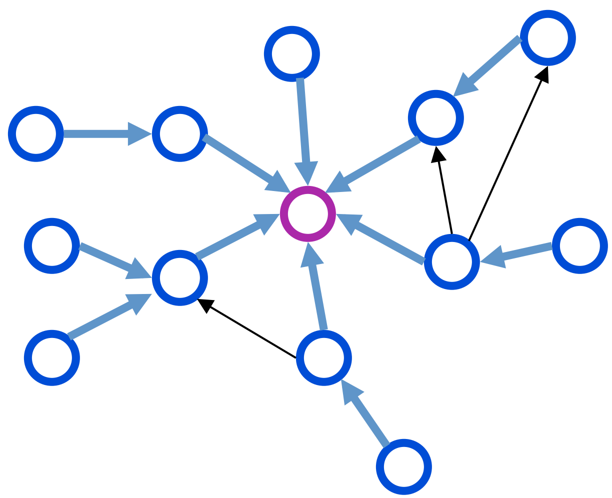




Extensions of undirected closeness centrality

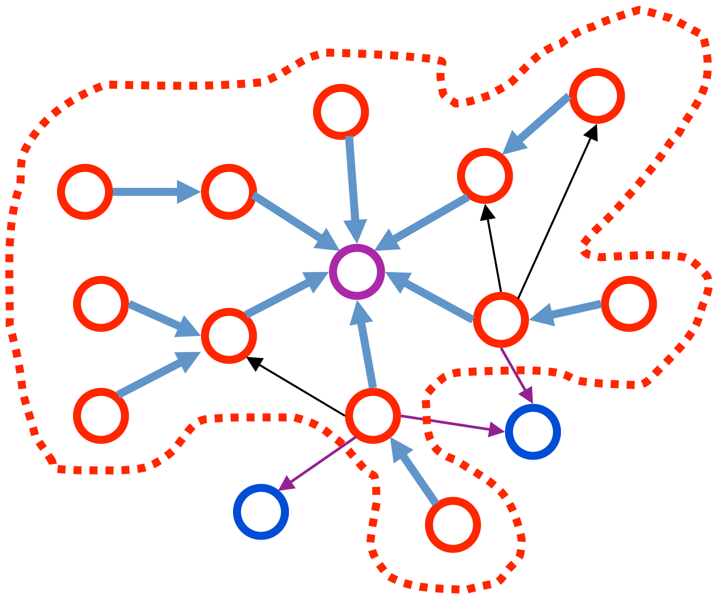
- closeness centrality usually implies
  - all paths should lead to you
  - paths should lead from you to everywhere else

- usually consider only vertices from which the node *i* in question can be reached



Influence range

- The influence range of *i* is the set of vertices who are reachable from the node *i*



Extending betweenness centrality to directed networks

- We now consider the fraction of all directed paths between any two vertices that pass through a node

betweenness of vertex i

paths between j and k that pass through i

$$C_B(i) = \sum_{j,k} g_{jk}(i) / g_{jk}$$

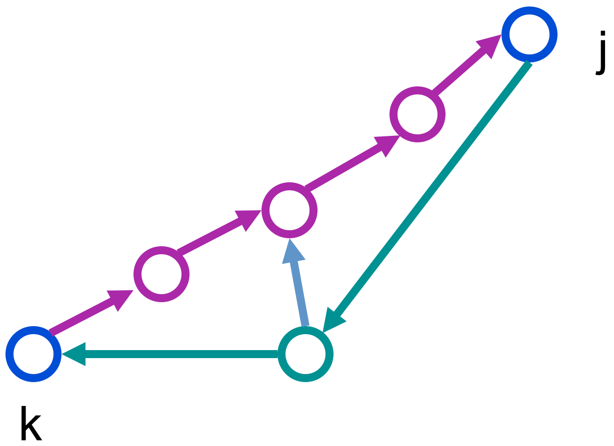
all paths between j and k

- Only modification: when normalizing, we have (N-1)*(N-2) instead of (N-1)*(N-2)/2, because we have twice as many ordered pairs as unordered pairs

$$C'_B(i) = C_B(i) / [(N-1)(N-2)]$$

Directed Shortest Path

- A node does not necessarily lie on a geodesic from *j* to *k* if it lies on a geodesic from *k* to *j*



Centrality based on neighborhood importance

- Node importance depends on the importance of its neighbors
- Let x_i be the importance of node I

$$x_i = \sum_{j=1}^{n} a_{ij} x_j$$

- How to calculate?

Also known as eigenvector centrality

- Iterative Process
- Initial values $x(0) = (x_1(0),\ldots,x_n(0))$
- Matrix form:

$$x(1)=A\,x(0)$$

- After t steps:

$$x(t)=A^t\,x(0)$$

- Eigenvector associated with eigenvalue:

$$A\,x=\kappa_1\,x$$

The most known Eigenvector Centrality Metric: PageRank

Motivation and introduction

- Why is Page Importance Rating important?
  - New challenges for information retrieval on the World Wide Web.
    - Huge number of web pages: 150 million by1998
      1000 billion by 2008
    - Diversity of web pages: different topics, different quality, etc.

- What is PageRank?
  - A method for rating the importance of web pages objectively and mechanically using the link structure of the web.

The history of PageRank

- PageRank was developed by Larry Page (hence the name *Page*-Rank) and Sergey Brin.

- It is first as part of a research project about a new kind of search engine. That project started in 1995 and led to a functional prototype in 1998.

- Shortly after, Page and Brin founded Google.
- 16 billion…

The history of PageRank

The Anatomy of a Large-Scale Hypertextual Web Search Engine

Sergey Brin and Lawrence Page
{sergey, page}@cs.stanford.edu
Computer Science Department, Stanford University, Stanford, CA 94305

Abstract

In this paper, we present Google, a prototype of a large-scale search engine which makes heavy use of the structure present in hypertext. Google is designed to crawl and index the Web efficiently and produce much more satisfying search results than existing systems. The prototype with a full text and hyperlink database of at least 24 million pages is available at http://google.stanford.edu/

To engineer a search engine is a challenging task. Search engines index tens to hundreds of millions of web pages involving a comparable number of distinct terms. They answer tens of millions of queries every day. Despite the importance of large-scale search engines on the web, very little academic research has been done on them. Furthermore, due to rapid advance in technology and web proliferation, creating a web search engine today is very different from three years ago. This paper provides an in-depth description of our large-scale web search engine -- the first such detailed public description we know of to date.

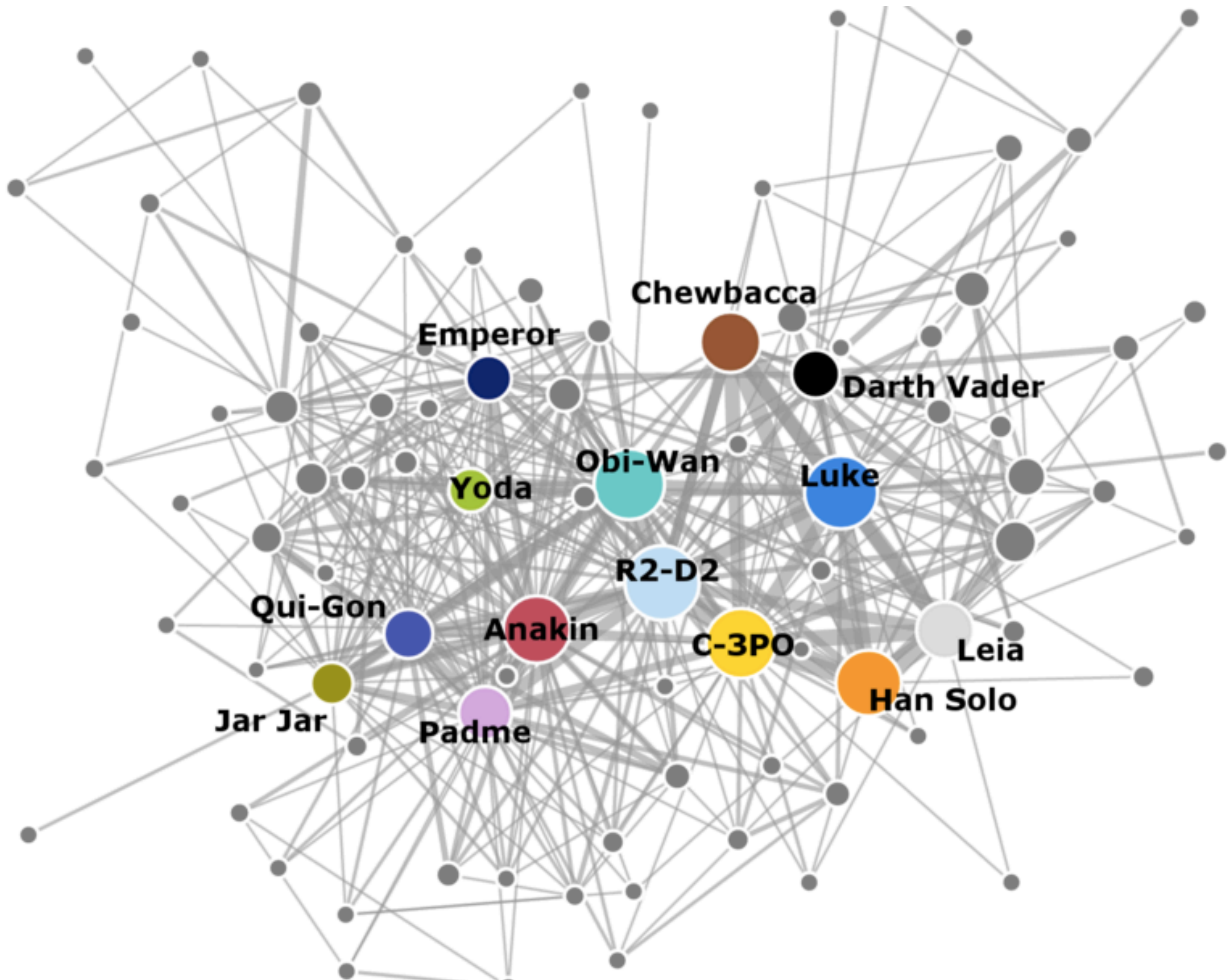
Apart from the problems of scaling traditional search techniques to data of this magnitude, there are new technical challenges involved with using the additional information present in hypertext to produce better search results. This paper addresses this question of how to build a practical large-scale system which can exploit the additional information present in hypertext. Also we look at the problem of how to effectively deal with uncontrolled hypertext collections where anyone can publish anything they want.

Keywords: World Wide Web, Search Engines, Information Retrieval, PageRank, Google

Some interesting references

1. J. Zhang, M. S. Ackerman, L. Adamic, Expertise Networks in Online Communities: Structure and Algorithms, WWW Conference 2007.
2. S. Brin, L. Page, The Anatomy of a Large-Scale Hypertextual Web Search Engine, WWW Conference 1999
3. A. Flores, A. Vieira, A. P. C. Silva, A. Ziviani, Fast Centraliy-Driven Diffusion in Dynamic Networks. Simplex, WWW, 2013.
4. E. Costa, A. Vieira, K. Wehmuth, A. Ziviani, A.P.C.Silva, Time Centrality in Dynamic Complex Networks, Advances Complex Systems, 2015
5. M. Cha, H. Haddadi, F. Benevenuto, K. Gummadi, Measuring User Influence in Twitter: The Million Follower Fallacy,2010
6. F. Riquelme, Measuring user influence on Twitter: A survey, 2015
7. M. Franceschet, PageRank: Standing on the Shoulder of Giants, Communications of ACM, 2010
8. P. Bonacich, P. Lloyd, Eigenvector-like measures of centrality for asymmetric relations, Social Networks, 2001

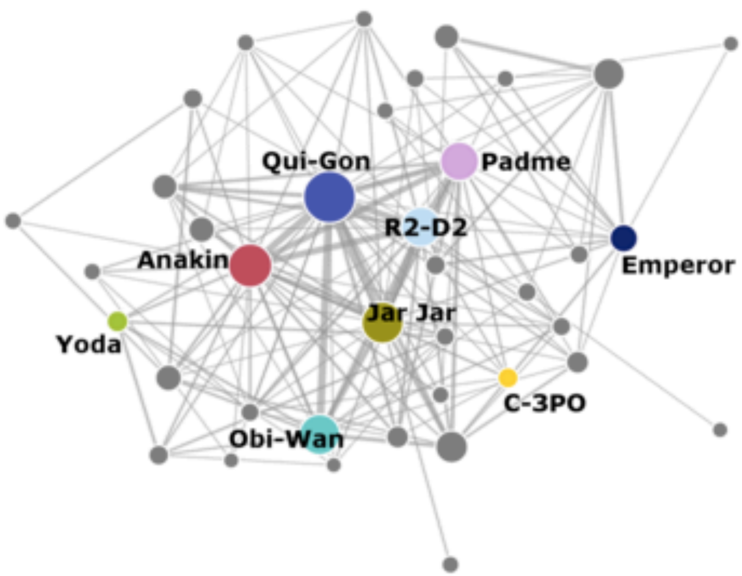
Star Wars network



http://evelinag.com/blog/2015/12-15-star-wars-social-network/index.html#.Vxu0BxJ94_W

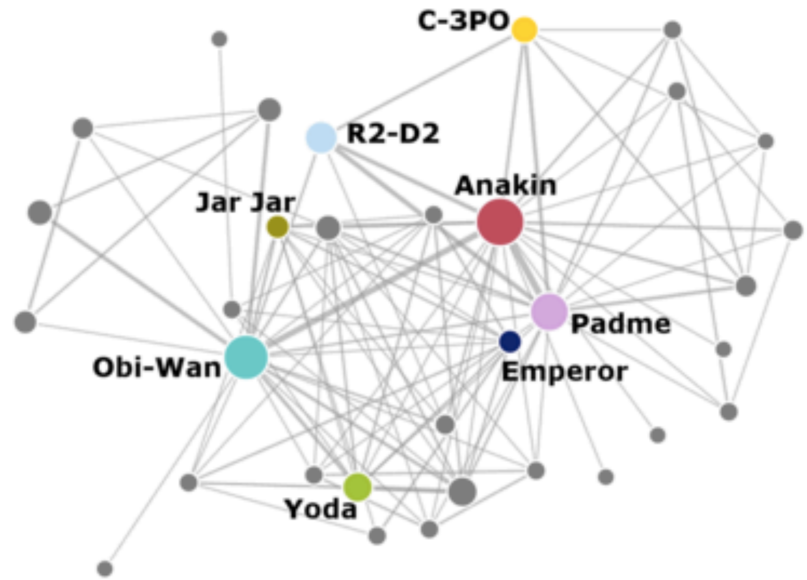
Star Wars network

Episode I: The Phantom Menace



Open network

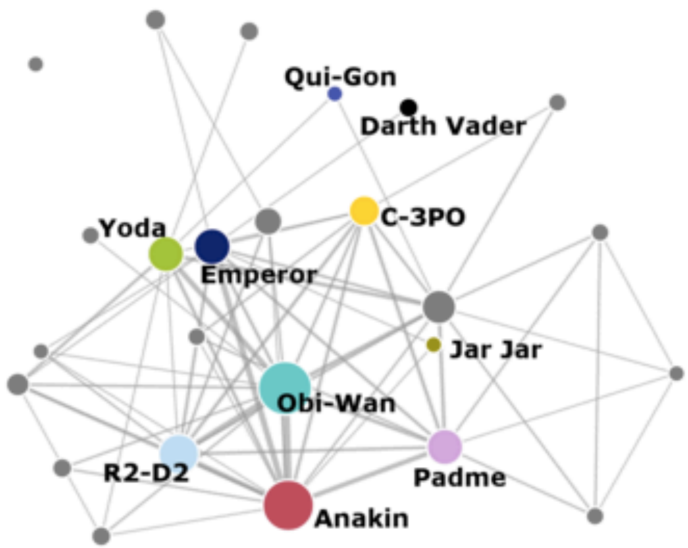
Episode II: Attack of the Clones



Open network

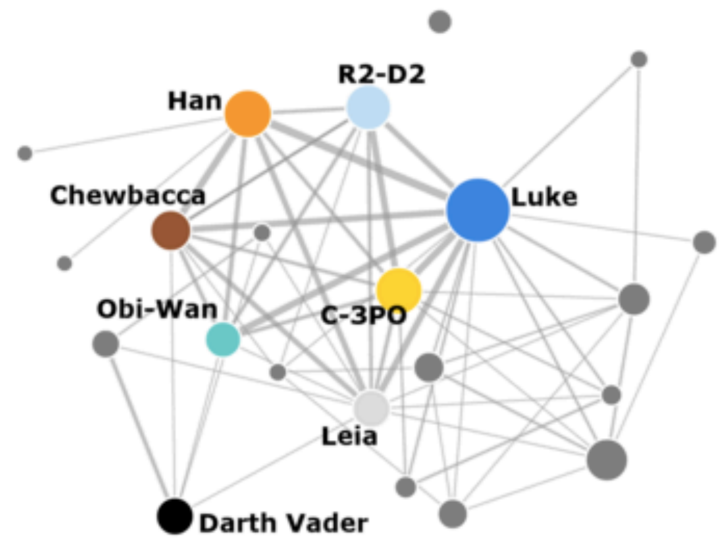
Star Wars network

Episode III: Revenge of the Sith



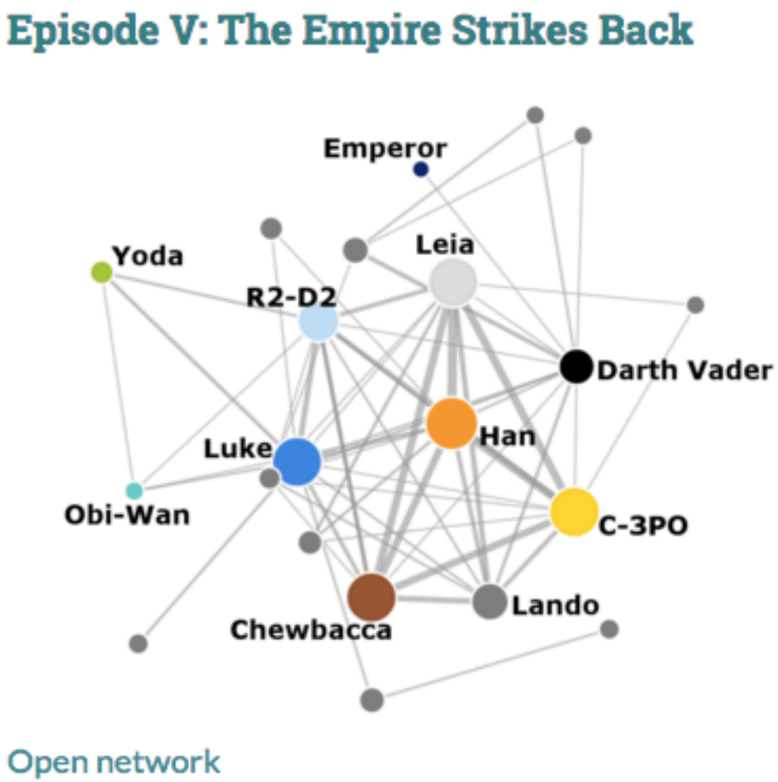
Open network

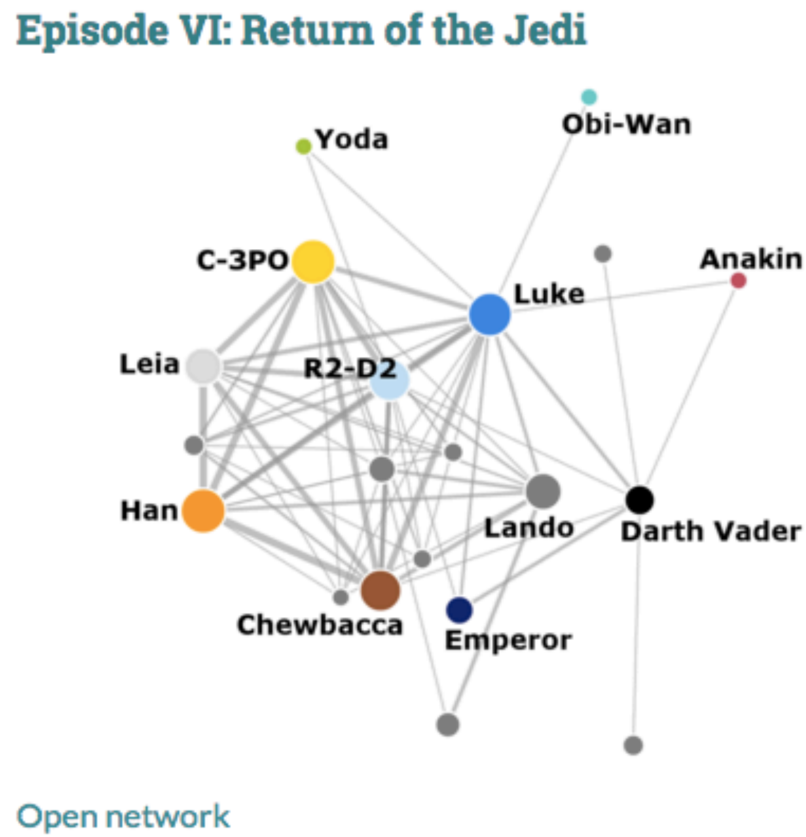
Episode IV: A New Hope



Open network

Star Wars network





Is Darth Vader/Anakin the central character?

- *Degree centrality* - this is simply the number of connections the node has in the network. In the Star Wars movies, this corresponds to the total number of scenes where each character speaks.

- *Betweenness* - this measure looks at how many shortest paths in the network lead through the node. For example, imagine you are Leia and you want to send a message to Greedo - the shortest path how to send it is via Han Solo, because he interacted both with Leia and with Greedo. On the other hand if you want to send a message to Luke, you don't have to go through Han because Leia knows Luke directly. The betweenness centrality for Han is computed using the number of shortest paths between all other characters that pass through him.

The two measures both show how important is a character in the network. The degree centrality shows how many people does each character interact with directly. The betweenness relates more to how integral each of the characters is to the story. Characters with high betweenness connect different areas of the social network.

Is Darth Vader/Anakin the central character?

Episode I

	Name	Degree
1.	QUI-GON	26
2.	ANAKIN	23
3.	JAR JAR	19
4.	R2-D2	19
5.	PADME	18

	Name	Betweenness
1.	QUI-GON	91.7
2.	JAR JAR	46.6
3.	EMPEROR	41.8
4.	R2-D2	30.9
5.	NUTE GUNRAY	27.2

Episode II

	Name	Degree
1.	ANAKIN	21
2.	OBI-WAN	19
3.	PADME	17
4.	YODA	10
5.	MACE WINDU	10

	Name	Betweenness
1.	OBI-WAN	64.7
2.	PADME	56.5
3.	MACE WINDU	12.7
4.	JAR JAR	8.3
5.	EMPEROR	6.8

Episode III

	Name	Degree
1.	ANAKIN	14
2.	OBI-WAN	13
3.	BAIL ORGANA	12
4.	EMPEROR	11
5.	PADME	10

	Name	Betweenness
1.	OBI-WAN	22.7
2.	EMPEROR	19.0
3.	PADME	8.0
4.	R2-D2	6.7
5.	BAIL ORGANA	4.5

Is Darth Vader/Anakin the central character?

It seems that Anakin is overall the most connected character in the first three films, based on his degree. He is however not very integral to the relations in the films! His betweenness score is so small he never makes it to the top-5 characters. This means that all the other characters interact directly between themselves rather than through Anakin. How do the same measures look for the original trilogy?

Is Darth Vader/Anakin the central character?

Episode IV

	Name	Degree
1.	LUKE	15
2.	LEIA	12
3.	C-3PO	10
4.	CHEWBACCA	9
5.	HAN	8

	Name	Betweenness
1.	LUKE	32.7
2.	LEIA	19.7
3.	HAN	15.0
4.	C-3PO	13.2
5.	CHEWBACCA	8.0

Episode V

	Name	Degree
1.	LUKE	12
2.	DARTH VADER	12
3.	HAN	11
4.	R2-D2	11
5.	C-3PO	10

	Name	Betweenness
1.	LUKE	25.2
2.	DARTH VADER	11.3
3.	LEIA	9.7
4.	HAN	6.7
5.	R2-D2	4.5

Episode VI

	Name	Degree
1.	LUKE	15
2.	R2-D2	12
3.	C-3PO	11
4.	LEIA	9
5.	HAN	9

	Name	Betweenness
1.	LUKE	24.3
2.	C-3PO	23.0
3.	DARTH VADER	18.5
4.	CHEWBACCA	16.0
5.	LANDO	5.5

Is Darth Vader/Anakin the central character?

Here both the centrality measures show very similar results - Luke is the most central character across all the films, and using both measures. The order of characters based on the two measures is almost the same.

The centrality analysis quantifies some of the things we could see from the social networks. The prequel trilogy has more complex social structures, with more interconnected characters. This also leads to the fact that Anakin is not that central to the story - some of the storylines happen alongside Anakin's story, or involve Anakin only on the side. On the other hand, the original trilogy has a more tight-knit structure. There is a smaller number of central characters and they bind the story together - this results into the agreement between the degree and betweenness centrality measures.

Is Darth Vader/Anakin the central character?

Perhaps this is part of the reason why the original trilogy is more popular - the plots are more consistent and driven by the main characters. The prequels have a more decentralized structure and no clear hero. Although the stories are linked by Anakin, he is not binding the other characters together.