Introduction to social network analysis

Russell J. Funk

Carlson School of Management
University of Minnesota

January 21, 2020

Roadmap

- Motivation and background
- Fundamentals
- Some real data
- Characterizing nodes
- Characterizing whole networks
- Software landscape
- Application: Disruptive inventions
- Next steps

Motivation and background

Why social network analysis?

Social networks matter for many things we care about...

- health and happiness (Bearman and Moody, 2004; Cornwell and Laumann, 2011)
- earnings and promotion (Burt, 1992; Mizruchi et al., 2011)
- employment and job search (Granovetter, 1995; Smith, 2005)

That includes science and technology, where networks matter for...

- creativity (Hargadon and Sutton, 1997; Fleming et al., 2007)
- achievement (Sekara et al., 2018; Coleman, 1988)
- careers (Azoulay et al., 2017; Whittington, 2018)

What is social network analysis?

Social network analysis consists of two distinct (but closely related) things. . .

A set of theories. . .

- for explaining why social networks matter for so many different outcomes.
- for explaining the origin of social network structures and how they change.

A set of methods

- for characterizing the properties of social networks along important dimensions.
- for relating those properties to various outcomes of interest.
- for modeling the dynamics of social networks over time.

Today, we'll be focused on the methodology of social network analysis

Where was social network analysis developed?

- Social network analysis (SNA) is a family of techniques for modeling relational data.
- But nothing about these techniques limits them to social networks.
- Consequently, development of SNA has been very interdisciplinary.

Historically, major contributions to SNA have come from...

- sociology (e.g., community, job search)
- anthropology (e.g., kinship)
- mathematics (e.g., graph theory, topology)
- physics (e.g., complexity)

But SNA is widely used in many other fields too...

- literature (e.g., text analysis)
- economics (e.g., education, development)
- health care (e.g., social determinants)
- social psychology (e.g., groups)
- marketing (e.g. influencers)
- computer science (e.g., information networks)

Fundamentals

Building blocks of networks

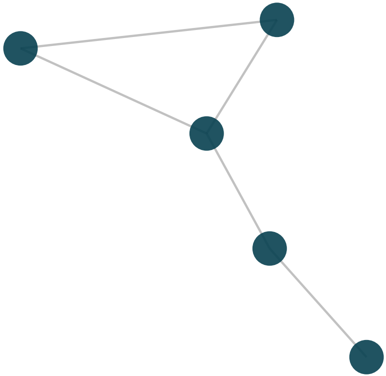
Networks are composed of two basic elements

Nodes

- A set of points.
- Also referred to as vertices.
- **Examples**—people, documents, words, ideas, websites.

Edges

- A set of relationships.
- Also referred to as ties.
- **Examples**—friendships, citations, co-occurrences, similarities, hyperlinks.



From building blocks to complex structures

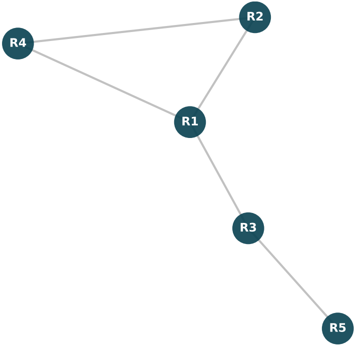
Overview of network data

- There are several different approaches for structuring network data.
- We'll discuss three of the most common.
  - Adjacency matrices
  - Edge lists
  - Affiliation matrices
- Each approach has its own strengths and weaknesses.
- Which is most appropriate will depend on the nature of your data.

Adjacency matrices

- Adjacency matrices represent relational data as an $n \times n$ matrix, where n is the number of nodes in the network.
- Each node appears once as a row and once as a column.
- Cell entries correspond to 0 or 1, indicating whether there is a relationship between node i and j.
- In the matrix on the right, we see connections among 5 nodes.
- Let's imagine these are researchers who are connected if they worked on a grant together.

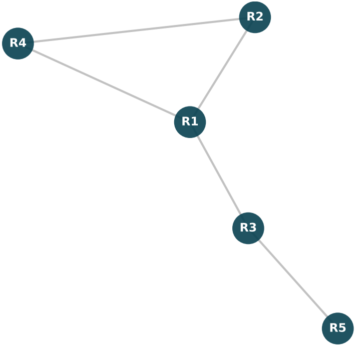
	R1	R2	R3	R4	R5
R1	0	-	-	-	-
R2	1	0	-	-	-
R3	1	0	0	-	-
R4	1	1	0	0	-
R5	0	0	1	0	0



Edge lists

source	target
R1	R2
R1	R3
R1	R4
R2	R4
R3	R5

- Edge lists essentially represent relational data in a sparse matrix format.
- Rather than recording all realized (1s) and potential (0s) relationships, we're only going to record those we observe.
- This format is really helpful for computational purposes when you have a large network.
- Note however that we lose information on isolates (nodes with no edges).



Affiliation matrices

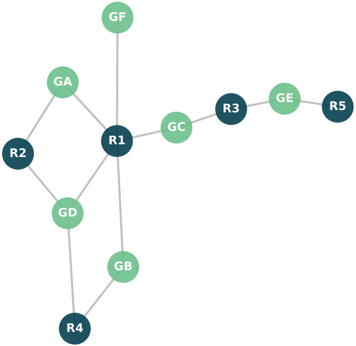
	GA	GB	GC	GD	GE	GF
R1	1	1	1	1	0	1
R2	1	0	0	1	0	0
R3	0	0	1	0	1	0
R4	0	1	0	1	0	0
R5	0	0	0	0	1	0

- Affiliation (or incidence) matrices are $n \times m$ matrices used to represent a network with two kinds of nodes.

- We'll call these nodes actors (n) and events (m), with rows indexing the former and columns indexing the latter.

- To continue our example, we can think of researchers as actors and grants as events.

- These kinds of networks are sometimes called "bipartite" or "two mode."

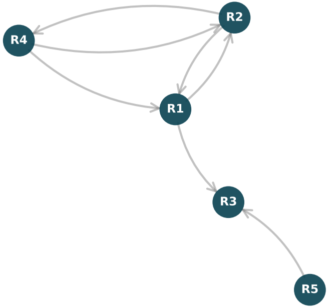


Augmenting the basics

Adding directionality to our edges

- Previously, we created as an undirected network; the edge A-B was identical to the edge B-A.
- But sometimes, we want to represent asymmetry in our edges.
- Consider citation networks, where papers published later in time can cite papers published earlier, but not vice versa.
- We can easily represent these kinds of relationships in our adjacency matrix by using cells above the diagonal.

	R1	R2	R3	R4	R5
R1	0	1	1	0	0
R2	1	0	0	1	0
R3	0	0	0	0	0
R4	1	1	0	0	0
R5	0	0	1	0	0

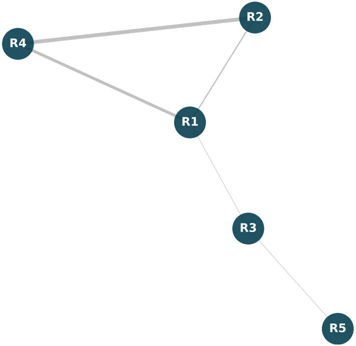


Augmenting the basics

Adding weight to our edges

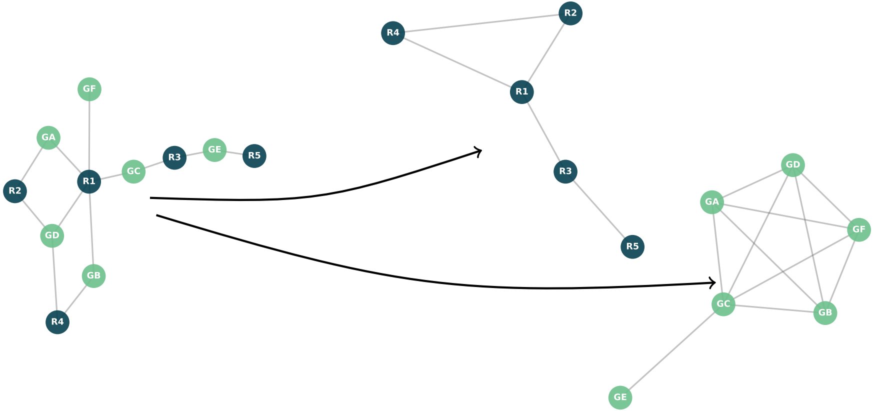
	R1	R2	R3	R4	R5
R1	0.0	-	-	-	-
R2	0.3	0.0	-	-	-
R3	0.1	0.0	0.0	-	-
R4	0.7	0.9	0.0	0.0	-
R5	0.0	0.0	0.1	0.0	0.0

- So far, we've considered edges in binary terms; they either exist or not, with nothing in between.
- But most relationships exist on a continuum; some are stronger (friends), some weaker (acquaintances)
- In social network analysis, we capture these differences by giving edges weights.
- We can easily represent such weights by using values other than 0 and 1 in our affiliation matrix.



Projecting an affiliation matrix

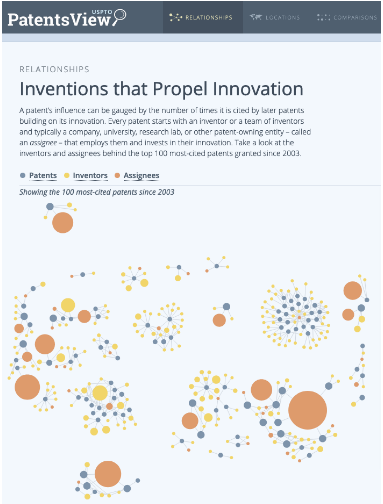
- Recall that in an affiliation matrix, we actually have two types of nodes.
- Intuitively, however, we tend to think about connections among people, rather than people being connected via ”events” (e.g., collaborations on grants).
- Therefore, we'll often want to project a bipartite network to a unipartite representation.



Some real data

Patent collaboration networks

- The examples we've been looking are a bit artificial.
- Our explorations might feel more meaningful with real data.
- Fortunately, it's easy to get network data using PatentsView, a researcher-focused interface to the USPTO.
- We'll focus specifically on patent collaborations among BTAA researchers.
- Two researchers are connected if they have collaborated together on a patent.



RELATIONSHIPS

Inventions that Propel Innovation

A patent's influence can be gauged by the number of times it is cited by later patents building on its innovation. Every patent starts with an inventor or a team of inventors and typically a company, university, research lab, or other patent-owning entity – called an *assignee* – that employs them and invests in their innovation. Take a look at the inventors and assignees behind the top 100 most-cited patents granted since 2003.

● Patents ● Inventors ● Assignees

Showing the 100 most-cited patents since 2003

(12) **United States Patent**
Kopelman et al.

(54) **OPTICAL SENSORS FOR THE DETECTION OF NITRIC OXIDE**

(75) Inventors: **Raoul Kopelman**, Ann Arbor, MI (US); **Heather Clark**, Middletown, CT (US); **Susan Barker**, Silver Spring, MD (US)

(73) Assignee: **The Regents of the University of Michigan**, Ann Arbor, MI (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

(21) Appl. No.: **09/365,487**

(22) Filed: **Aug. 2, 1999**

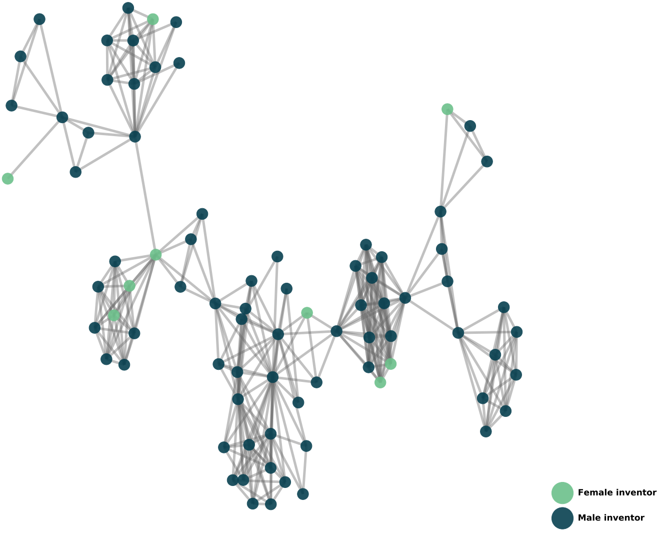
(51) **Int. Cl.**[7] **G02B 6/00**

(52) **U.S. Cl.** **385/12**; 385/38; 385/127; 385/128

(58) **Field of Search** 385/12, 33, 38, 385/43, 123, 127, 128, 141–145; 359/368, 385; 250/227.14, 227.11; 427/162, 163, 457, 487; 435/68, 4, 176, 6, 91.2; 536/22.1, 23.1, 24.3, 24.5

University of Michigan

Main component only, 2007–2012

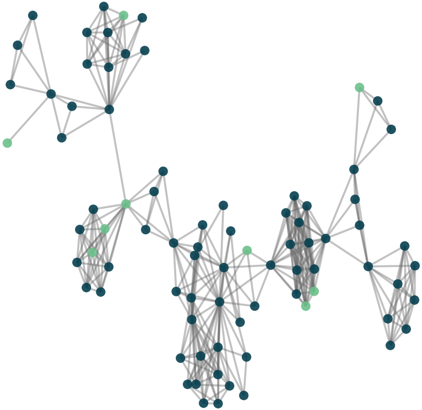


Characterizing nodes

Characterizing nodes

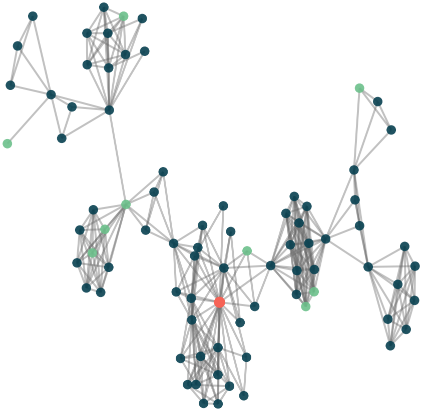
So you've got a network...

- One of the most common things we'll want to do with these sorts of data is look at differences among nodes.
- For example, we might want to characterize the relative importance of different inventors in the network.
- Or perhaps we want to examine differences in how various inventors are connected to others.
- Let's look at a handful of useful node-level metrics.



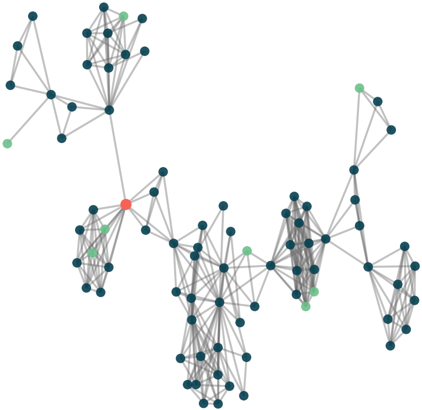
Degree centrality

- Degree captures the number of nodes to which the focal node is connected.
- Sometimes the measure is normalized by the number of nodes in the network.
- Despite its simplicity, the measure works really well for many purposes.
- The red node in the network to the right has the highest degree centrality.



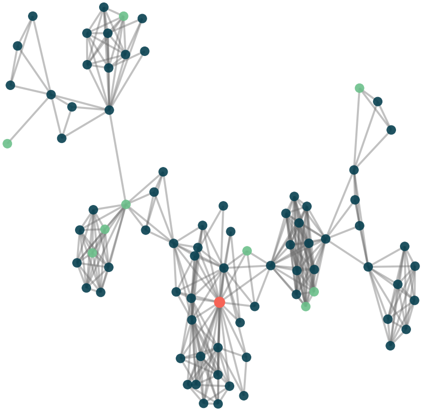
Betweenness centrality

- Betweenness captures the number of shortest paths between pairs of nodes that pass through the focal node.
- Sometimes the measure is normalized by the number of shortest paths between pairs of nodes in the network.
- Typically, people with higher betweenness centrality have better access to information.
- The red node in the network to the right has the highest betweenness centrality.

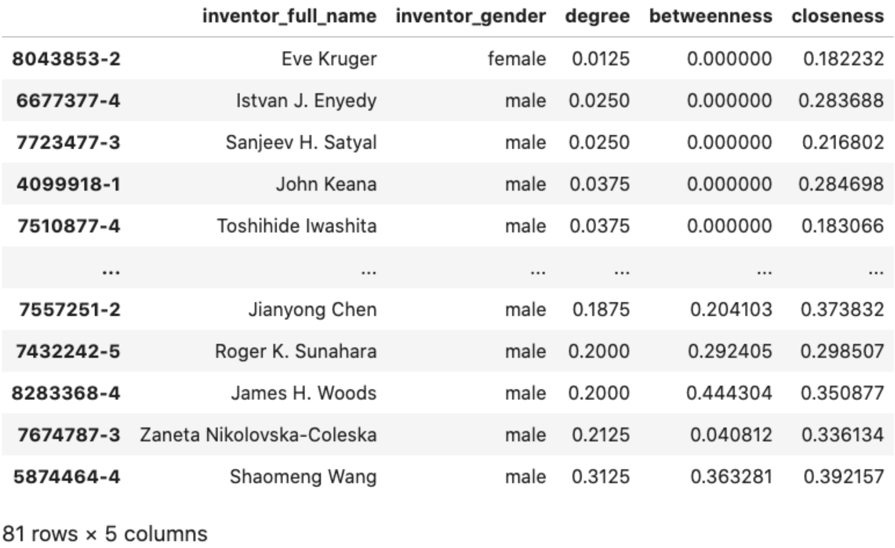


Closeness centrality

- Closeness captures the inverse of the sum of the length of the shortest paths from the focal node to others in the network.
- Sometimes the measure is normalized by using the average length of the shortest paths (rather than the sum).
- The red node in the network to the right has the highest closeness centrality.
- Notice that the inventor with the highest closeness is also the inventor who had the highest degree.

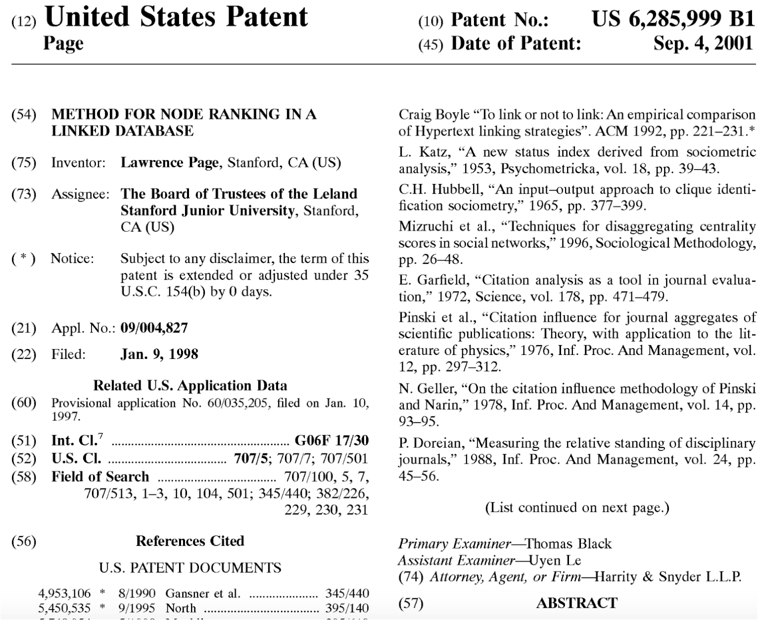


High and low centralities



	inventor_full_name	inventor_gender	degree	betweenness	closeness
8043853-2	Eve Kruger	female	0.0125	0.000000	0.182232
6677377-4	Istvan J. Enyedy	male	0.0250	0.000000	0.283688
7723477-3	Sanjeev H. Satyal	male	0.0250	0.000000	0.216802
4099918-1	John Keana	male	0.0375	0.000000	0.284698
7510877-4	Toshihide Iwashita	male	0.0375	0.000000	0.183066
...	...	...	...	...	...
7557251-2	Jianyong Chen	male	0.1875	0.204103	0.373832
7432242-5	Roger K. Sunahara	male	0.2000	0.292405	0.298507
8283368-4	James H. Woods	male	0.2000	0.444304	0.350877
7674787-3	Zaneta Nikolovska-Coleska	male	0.2125	0.040812	0.336134
5874464-4	Shaomeng Wang	male	0.3125	0.363281	0.392157

81 rows × 5 columns



(12) **United States Patent**
Page

(10) **Patent No.: US 6,285,999 B1**
(45) **Date of Patent: Sep. 4, 2001**

(54) **METHOD FOR NODE RANKING IN A LINKED DATABASE**

(75) Inventor: **Lawrence Page**, Stanford, CA (US)

(73) Assignee: **The Board of Trustees of the Leland Stanford Junior University**, Stanford, CA (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

(21) Appl. No.: **09/004,827**

(22) Filed: **Jan. 9, 1998**

Related U.S. Application Data

(60) Provisional application No. 60/035,205, filed on Jan. 10, 1997.

(51) **Int. Cl.**[7] **G06F 17/30**
(52) **U.S. Cl.** **707/5**; 707/7; 707/501
(58) **Field of Search** 707/100, 5, 7, 707/513, 1–3, 10, 104, 501; 345/440; 382/226, 229, 230, 231

(56) **References Cited**

U.S. PATENT DOCUMENTS

4,953,106	*	8/1990	Gansner et al.	345/440
5,450,535	*	9/1995	North	395/140

Craig Boyle "To link or not to link: An empirical comparison of Hypertext linking strategies". ACM 1992, pp. 221–231.*

L. Katz, "A new status index derived from sociometric analysis," 1953, Psychometricka, vol. 18, pp. 39–43.

C.H. Hubbell, "An input–output approach to clique identification sociometry," 1965, pp. 377–399.

Mizruchi et al., "Techniques for disaggregating centrality scores in social networks," 1996, Sociological Methodology, pp. 26–48.

E. Garfield, "Citation analysis as a tool in journal evaluation," 1972, Science, vol. 178, pp. 471–479.

Pinski et al., "Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics," 1976, Inf. Proc. And Management, vol. 12, pp. 297–312.

N. Geller, "On the citation influence methodology of Pinski and Narin," 1978, Inf. Proc. And Management, vol. 14, pp. 93–95.

P. Doreian, "Measuring the relative standing of disciplinary journals," 1988, Inf. Proc. And Management, vol. 24, pp. 45–56.

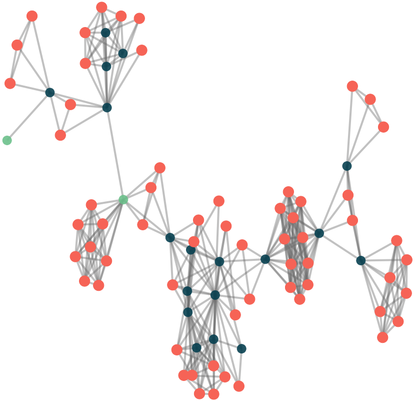
(List continued on next page.)

Primary Examiner—Thomas Black
Assistant Examiner—Uyen Le
(74) *Attorney, Agent, or Firm*—Harrity & Snyder L.L.P.

(57) **ABSTRACT**

Clustering

- Clustering captures the degree to which we observe that if A is connected to B and C then B and C are also connected.
- This phenomenon is sometimes known as triadic closure.
- High clustering is one of the most distinctive features of social networks.
- We often see very high clustering in projections of bipartite networks.
- The red nodes in the network all have clustering values of 1.0.



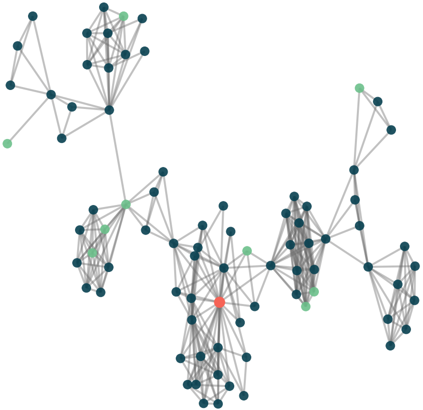
Constraint

- Constraint (essentially) captures the degree of interconnection among a node's contacts.

- Nodes that have low constraint are thought to occupy "structural holes."

- Thus, we are often interested in nodes with low rather than high constraint.

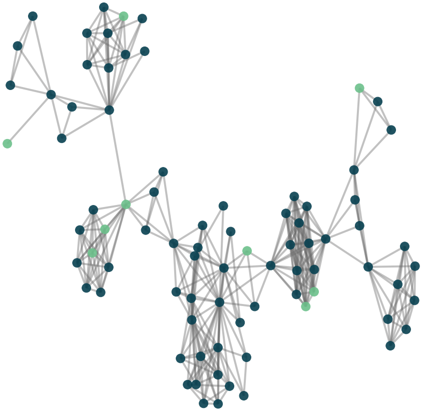
- The red node in the network to the right has the lowest constraint.



Characterizing whole networks

Characterizing whole networks

- There are also many measures available for characterizing whole networks.
- These measures give us insight into overall patterns of connectivity.
- As with the node level measures, there are a few big categories of whole network measures you'll likely encounter.



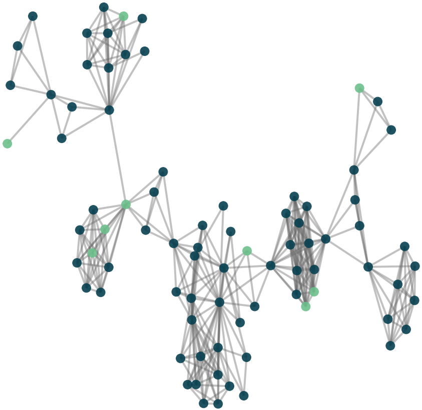
Density

- Density captures the number of edges observed in the network relative to the number of possible edges.

- This measure is probably both the simplest and most common whole network measures you'll encounter.

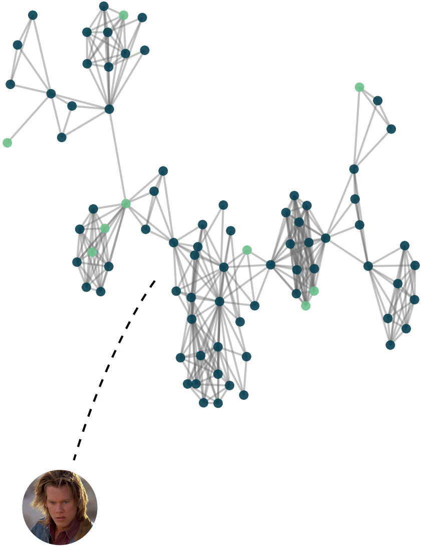
- In the Michigan collaboration network, the average path length is **0.43**.

- Thus, you could say that the network is 43% connected.



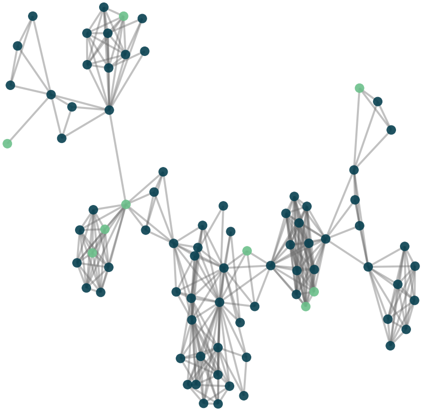
Average (shortest) path length

- Average (shortest) path length captures the expected number of edges between two randomly chosen nodes in a network.
- If you’re familiar with the “small-world” phenomenon or the “Kevin Bacon game,” then you know about this measure.
- Social networks typically have lower path lengths than we would think intuitively (c.f., “six degrees of separation).
- In the Michigan collaboration network, the average path length is **1.60**.



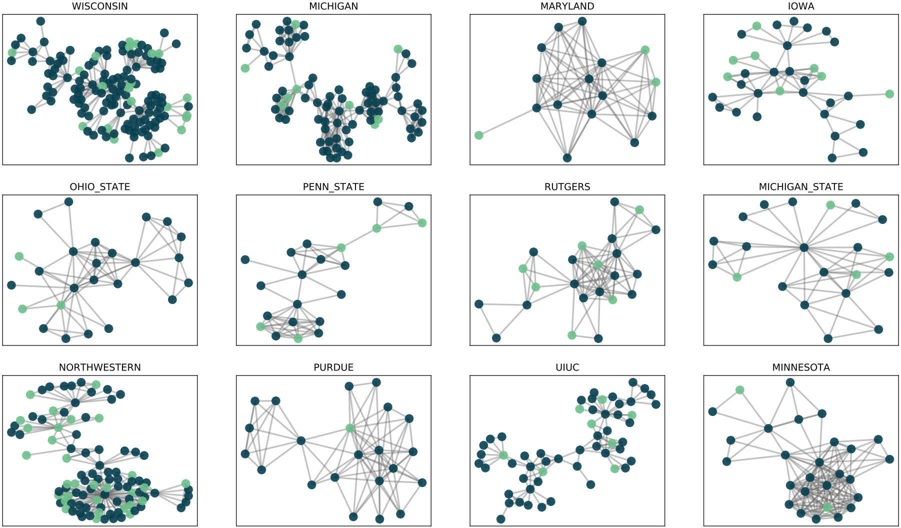
Clustering

- We can measure clustering at the whole network level similar to how we can at the node level.
- Together with average path length, clustering is one of the defining properties of small-world networks.
- Small world networks tend to have high clustering and low average path lengths.
- In the Michigan collaboration network, the clustering coefficient is **0.67**.



BTAA inventor networks (sans Nebraska and Indiana)

Main component only, 2007–2012



WISCONSIN

MICHIGAN

MARYLAND

IOWA

OHIO_STATE

PENN_STATE

RUTGERS

MICHIGAN_STATE

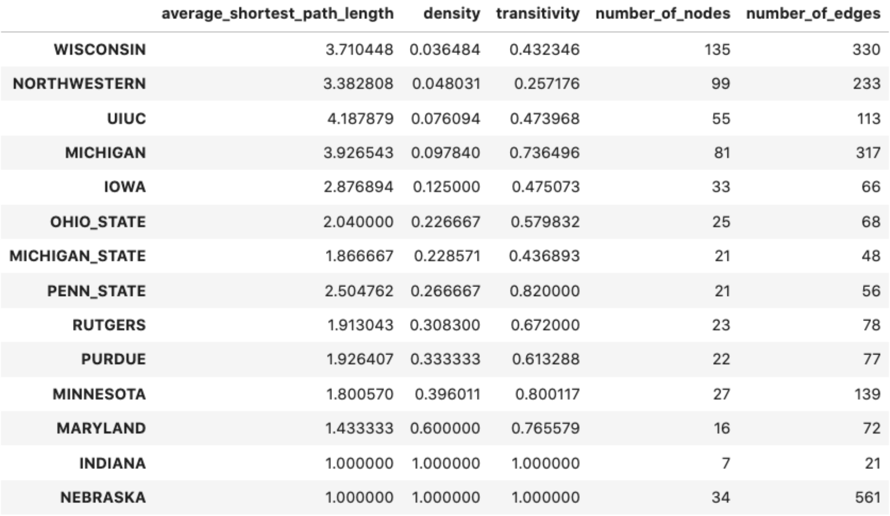
NORTHWESTERN

PURDUE

UIUC

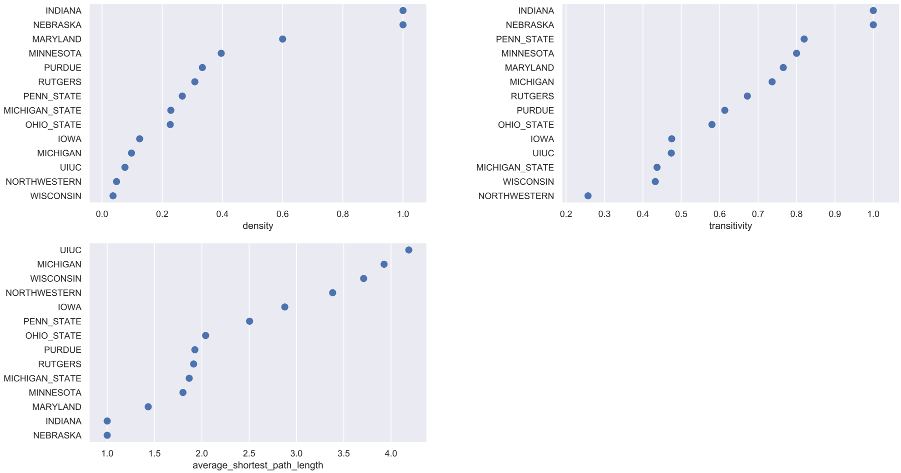
MINNESOTA

Comparing whole networks



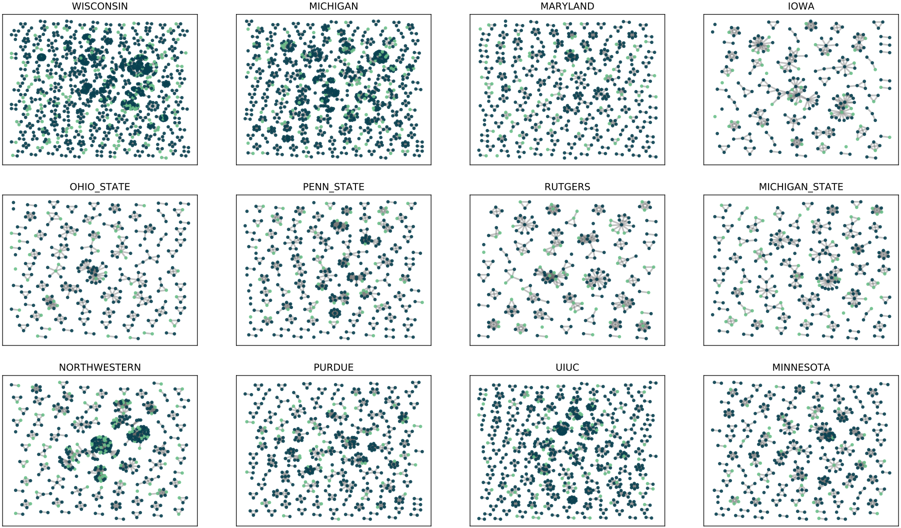
	average_shortest_path_length	density	transitivity	number_of_nodes	number_of_edges
WISCONSIN	3.710448	0.036484	0.432346	135	330
NORTHWESTERN	3.382808	0.048031	0.257176	99	233
UIUC	4.187879	0.076094	0.473968	55	113
MICHIGAN	3.926543	0.097840	0.736496	81	317
IOWA	2.876894	0.125000	0.475073	33	66
OHIO_STATE	2.040000	0.226667	0.579832	25	68
MICHIGAN_STATE	1.866667	0.228571	0.436893	21	48
PENN_STATE	2.504762	0.266667	0.820000	21	56
RUTGERS	1.913043	0.308300	0.672000	23	78
PURDUE	1.926407	0.333333	0.613288	22	77
MINNESOTA	1.800570	0.396011	0.800117	27	139
MARYLAND	1.433333	0.600000	0.765579	16	72
INDIANA	1.000000	1.000000	1.000000	7	21
NEBRASKA	1.000000	1.000000	1.000000	34	561

Comparing whole networks



BTAA inventor networks (sans Nebraska and Indiana)

Full network, 2007–2012



Software landscape

Python!

Community

- Python is one of the most popular programming languages.
- That means there's a big community of developers who are regularly creating both general purpose and highly specialized packages for network analysis.
- There are also many people who can help with questions (e.g., via Stack Overflow).

Flexibility

- Because Python is a general-purpose programming language, it's easy to write programs to do things that are not implemented in existing libraries.
- In addition, as you learn network analysis in Python, you'll also be learning about Python generally, meaning you'll be able to use it for things beyond network analysis.

Packages

The two most popular packages are igraph and networkx.

igraph

- Pros
  - Ultra fast; implemented in C
  - Has companion libraries in R and C
  - Great suite of community detection algorithms
- Cons
  - Can be difficult to install
  - Development has stalled (the last major release was in 2015)

networkx

- Pros
  - Large suite of network measures
  - Integrates easily with matplotlib
  - Development is active
- Cons
  - Slower; implemented in pure Python

Beyond Python

Notwithstanding its strengths, Python's not the only game in town.

R

- statnet—Tools for the statistical analysis of networks (e.g., ERGMs).
- RSiena—Also focuses on the statistical analysis of networks.
- igraph—The same igraph you know and love from Python, but in R.

Pajek

- Slovenian for "spider," Pajek is a popular, GUI based program for Windows.
- Primarily used for visualization.

UCINET

- Like Pajek, UCINET is a popular, GUI based network analysis program for Windows.
- Less emphasis on visualization and more focus on quantitative analysis.

Gephi

- Gephi is a popular, GUI based program for large scale network visualization.

Cytoscape

- Similar to Gephi, Cytoscape is a popular GUI based program for network visualization.
- There's also a package available that allows you to use Cytoscape in Python scripts.

Application: Disruptive inventions

From social network analysis to substantive insights

- In the last few minutes, I want to show you how social network analysis can lead to new substantive insights.
- So far, our examples have used network analysis to look at relationships among people.
- But as I've mentioned, nothing limits our use of network analysis to social networks.
- In this example, we'll apply social network analysis to examine relationships among documents.
- We'll see how we can use network analysis to find disruptive papers and patents.

Conceptual motivation

Foundational theories of innovation distinguish between two types of discoveries.

- The **first** type builds on and therefore **consolidates** existing knowledge.
  - For example, glyphosate resistant soybeans made the herbicide glyphosate more valuable and therefore reinforced its widespread use.
- The **second** type departs from and therefore **destabilizes** existing knowledge.
  - For example, recombinant DNA was destabilizing because it helped introduce a fundamentally new method of drug discovery.

Existing measures

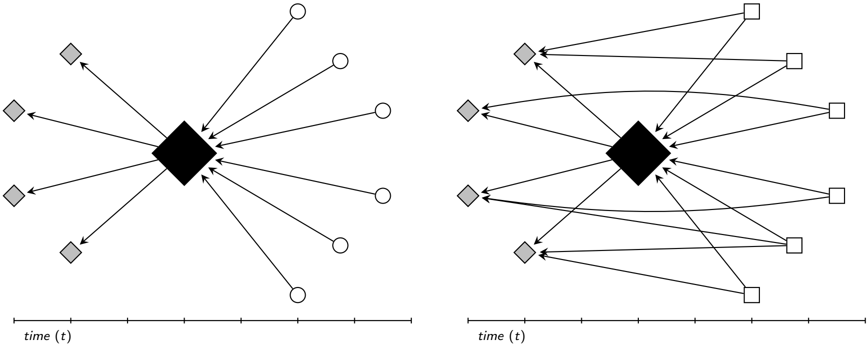
However, established bibliometric measures do not capture this distinction.

Instead, most measures fall into one of two categories.

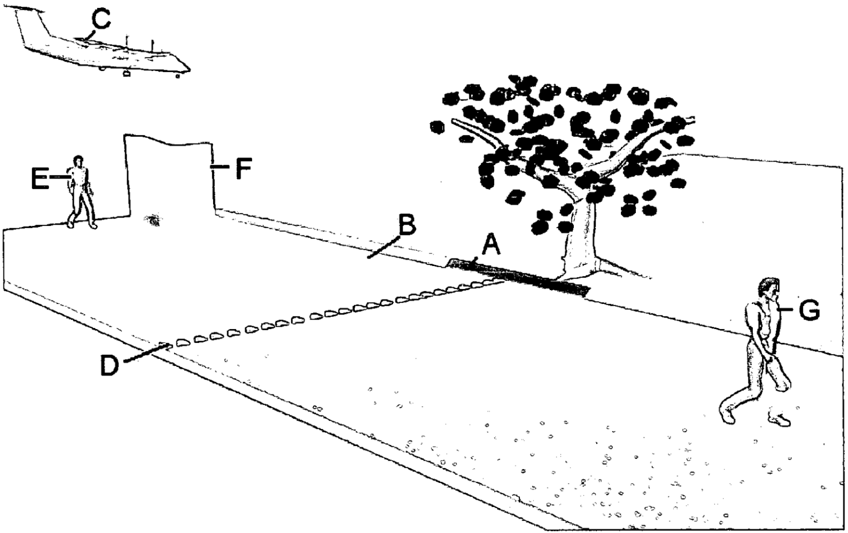
- First, some measures focus on quantifying the extent to which an idea is used.
  - Examples include measures that count citations to patents or papers.
- Second, some measures focus on quantifying the extent to which an idea is distinctive from the status quo.
  - Examples include measures that identify papers or patents that encompass previously unconnected keywords or technology areas.

Although often informative, these approaches have some limitations. . .

Let me focus on the more widely used category first



Using counts of forward citations (impact) we could not differentiate these two papers.



C
E
F
B
A
G
D

(19) **United States**

(12) **Patent Application Publication**

St. Clair

(10) **Pub. No.: US 2006/0071122 A1**

(43) **Pub. Date: Apr. 6, 2006**

(54) **FULL BODY TELEPORTATION SYSTEM**

(76) Inventor: **John Quincy St. Clair**, San Juan, PR (US)

Correspondence Address:
JOHN ST. CLAIR
4A
52 KINGS COURT
SAN JUAN, PR 00911 (US)

(21) Appl. No.: **10/953,212**

(22) Filed: **Sep. 29, 2004**

Publication Classification

(51) **Int. Cl.**
B64C 39/00 (2006.01)

(52) **U.S. Cl.** **244/62**

(57) **ABSTRACT**

A pulsed gravitational wave wormhole generator system that teleports a human being through hyperspace from one location to another.

BACKGROUND OF THE INVENTION

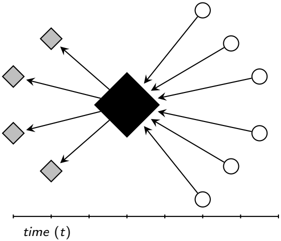
[0002] The basis for this invention is an event, referring to **FIG. 1**, occurring on May 2, 2004, in which the inventor ("he") personally experienced a full-body teleportation while walking to the bus stop (A) along a road (B) that runs perpendicular to the nearby commercial airport runways where planes are landing. There is a wide iron grating (D) for water drainage that crosses the road at the center of the bus stop. The grating width is such that one has to make a concerted effort to jump across it in order to get from one side to the other. Approximately 50 meters from the iron grating, he (E) felt a vertical wave (F), similar to a flag waving in the breeze, traveling down the street toward the bus stop. The wave velocity was about 1 meter per second, which was slightly faster than his walking speed. In the next instance, he (G) found himself down the street near the corner of the next block. Realizing that he had passed the bus stop, he turned around to see the iron grating approximately 50 meters up the street in back of him. Because there was no recollection of having jumped across the iron grating nor of having passed the bus stop's yellow marker line, he realized that he had been teleported a distance of 100 meters while moving along with the traveling wave. It was obvious that the wave was pulsed because the front edge overtook the

[0003] It took a number of days in order to understand this sequence of events. The explanation involves knowledge of a wide range of subjects such as gravitation physics, hyperspace physics, wormhole electromagnetic theory and experimentation, quantum physics, and the nature of the human energy field.

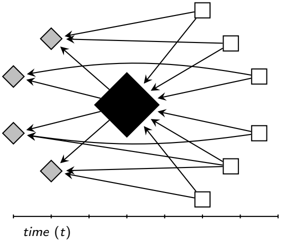
[0004] It is obvious from the above scenario that the airplane momentarily crossing perpendicular to the road generates the aforementioned pulse. Because the airplane has an engine on each wing, there are two propellers which conceivably are rotating out-of-phase with each other. That is, the blade of one propeller could be pointing up and the equivalent blade on the other engine could be pointing in a slightly different direction. Notice that the tip of the blade traces out a helix as the plane is landing.

What do we propose?

- The measure of an intellectual contribution is how it influences existing knowledge.
- We consider an intellectual contribution to be. . .
  - **consolidating** when it increases the use of its predecessors.
  - **destabilizing** when it decreases the use of its predecessors.



Maximally destabilizing



Maximally consolidating

Some math

We distinguish between three 'types' of documents that can be cited by subsequent documents

1. the focal document, f, which we seek to assess
2. a set of predecessors cited by the focal document, b
3. a set of forward citations, i

Let

$$f_{it} = \begin{cases} 1 & \text{if } i \text{ cites the focal document (type } f\text{)} \\ 0 & \text{otherwise,} \end{cases}$$

and

$$b_{it} = \begin{cases} 1 & \text{if } i \text{ cites any focal document predecessors (type } b\text{)} \\ 0 & \text{otherwise,} \end{cases}$$

We then define our measure as

$$CD_t = \frac{1}{n_t} \sum_{i=1}^{n} \frac{-2 f_{it} b_{it} + f_{it}}{w_{it}}, \quad w_{it} > 0.$$

In a different measure, we weight by m_t (i.e., forward citations)...

$$mCD_t = \frac{m_t}{n_t} \sum_{i=1}^{n} \frac{-2 f_{it} b_{it} + f_{it}}{w_{it}}, \quad w_{it} > 0,$$

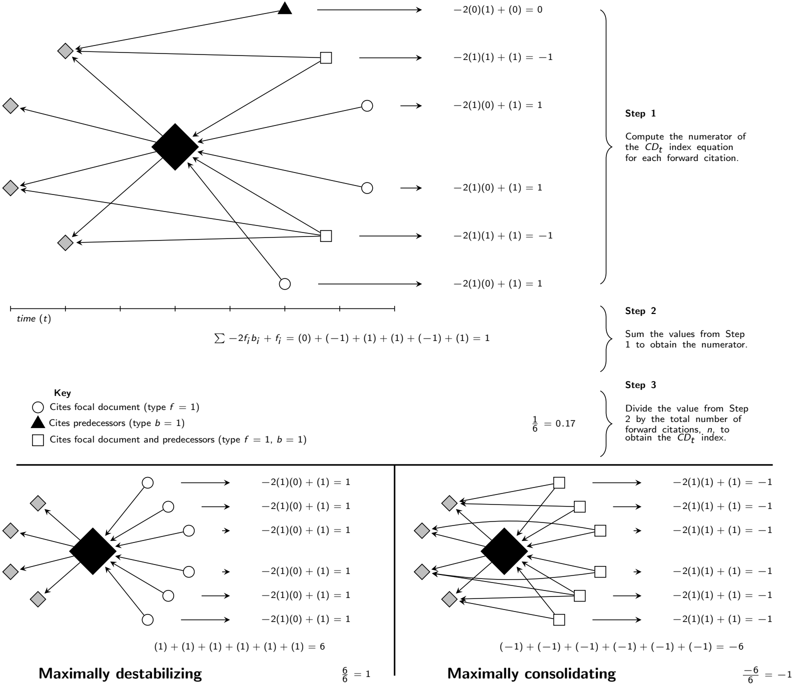
CD_t

- ranges from -1 to 1
- -1 $\longrightarrow$ consolidating
- 1 $\longrightarrow$ destabilizing

mCD_t

- - $\longrightarrow$ consolidating
- $+$ $\longrightarrow$ destabilizing
- forward cite weight m_t

...we usually set $t = 5$.



Step 1

Compute the numerator of the CD_t index equation for each forward citation.

Step 2

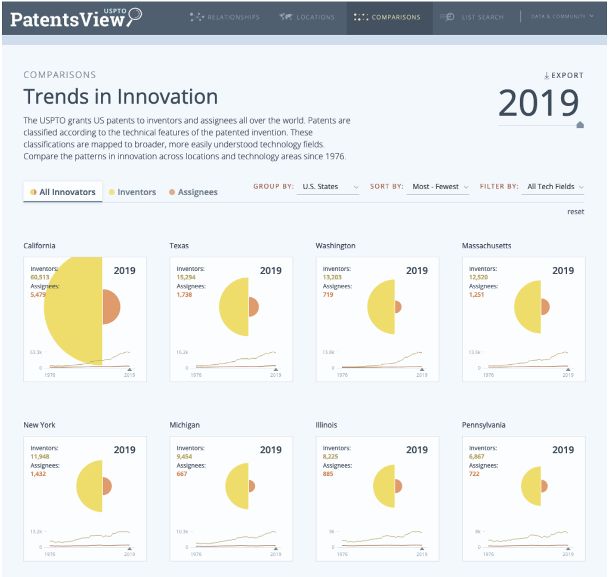
Sum the values from Step 1 to obtain the numerator.

Step 3

Divide the value from Step 2 by the total number of forward citations, n, to obtain the CD_t index.

Back to PatentsView

We'll apply the CD_t index to U.S. patents



COMPARISONS

↓EXPORT

Trends in Innovation

2019

The USPTO grants US patents to inventors and assignees all over the world. Patents are classified according to the technical features of the patented invention. These classifications are mapped to broader, more easily understood technology fields. Compare the patterns in innovation across locations and technology areas since 1976.

All Innovators · Inventors · Assignees

GROUP BY: U.S. States · SORT BY: Most - Fewest · FILTER BY: All Tech Fields

reset

California: 2019 · Inventors: 60,513 · Assignees: 5,479

Texas: 2019 · Inventors: 15,294 · Assignees: 1,738

Washington: 2019 · Inventors: 13,203 · Assignees: 719

Massachusetts: 2019 · Inventors: 12,520 · Assignees: 1,251

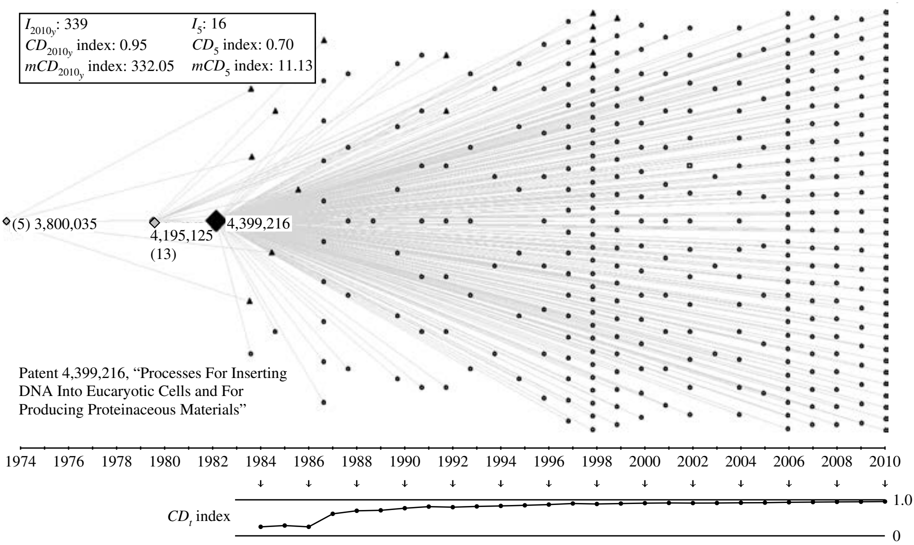
New York: 2019 · Inventors: 11,948 · Assignees: 1,432

Michigan: 2019 · Inventors: 9,454 · Assignees: 667

Illinois: 2019 · Inventors: 8,225 · Assignees: 885

Pennsylvania: 2019 · Inventors: 6,867 · Assignees: 722

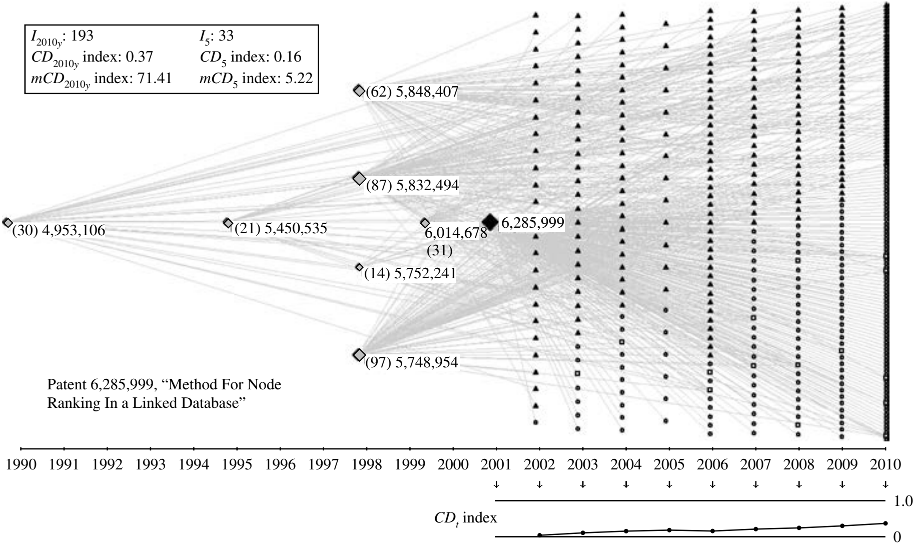
Highly destabilizing patent: Eukaryotic Cotransformation



I_{2010_y}: 339	I_5: 16
CD_{2010_y} index: 0.95	CD_5 index: 0.70
mCD_{2010_y} index: 332.05	mCD_5 index: 11.13

Patent 4,399,216, “Processes For Inserting DNA Into Eucaryotic Cells and For Producing Proteinaceous Materials”

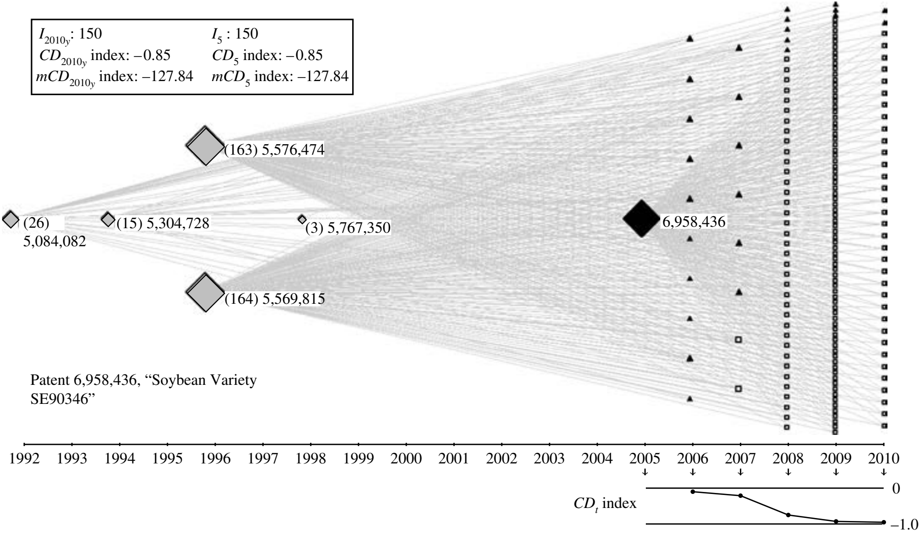
Moderately destabilizing patent: PageRank



I_{2010y}: 193 I_5: 33
CD_{2010y} index: 0.37 CD_5 index: 0.16
mCD_{2010y} index: 71.41 mCD_5 index: 5.22

Patent 6,285,999, “Method For Node
Ranking In a Linked Database”

Highly consolidating patent: Resistant soybeans



I_{2010y}: 150 I_5 : 150

CD_{2010y} index: –0.85 CD_5 index: –0.85

mCD_{2010y} index: –127.84 mCD_5 index: –127.84

Patent 6,958,436, “Soybean Variety SE90346”

Mini case: University research commercialization

Sample

- 110 US universities ever in the top 100 by federal R&D funding, 1981-2005
- USPTO utility patents ($\approx$ 45,000), 1993-2005 (Google Patents, USPTO)

Four dependent variables $(t+1)$

- impact
- new combinations (distinctiveness)
- disruptiveness
- radicalness

Independent variables

- scientific productivity and impact (ISI University Indicators)
- technology transfer office age (AUTM Licensing Survey)
- industrial and federal R&D funding (NSF WebCaspar)
- industry contractual ties (USKE project)

Estimation

- Two-way fixed effects

Models of university patenting, 1993-2005†

	Impact (I_t)	Combinations	CD_t	mCD_t
Commercial experience				
Patent stock	−0.0003	−0.0049	0.0000	-1.2697***
	(0.0010)	(0.0191)	(0.0003)	(0.1498)
TTage	0.2337***	6.5645**	-0.1304***	61.0185***
	(0.0589)	(2.0611)	(0.0324)	(16.1994)
TTage^2	-0.0225*	0.3851+	0.0012	-18.0159***
	(0.0099)	(0.2252)	(0.0035)	(1.7697)
Industry ties				
Industry R&D	0.0038*	0.0512+	−0.0000	-0.9919***
	(0.0019)	(0.0300)	(0.0005)	(0.2360)
Industry contracts	-0.0157*	-0.2666+	−0.0001	-6.7401***
	(0.0079)	(0.1478)	(0.0023)	(1.1615)
Government ties				
NSF grants	−0.0272	0.0807	−0.0039	11.0251*
	(0.0362)	(0.6408)	(0.0101)	(5.0366)
NIH grants	−0.0105	−0.9679	0.0346+	19.4857*
	(0.0568)	(1.1588)	(0.0182)	(9.1080)
Government interest	0.0143***	−0.0095	0.0002	0.6960**
	(0.0015)	(0.0307)	(0.0005)	(0.2410)
Scientific capacity				
Scientific articles	0.5075***	−1.9103	0.0398	−7.5977
	(0.1268)	(3.2813)	(0.0516)	(25.7895)
Impact factor	0.0303***	−0.0919	0.0031**	5.0206***
	(0.0046)	(0.0599)	(0.0009)	(0.4710)

†Models include additional controls

LETTER

https://doi.org/10.1038/s41586-019-0941-9

Large teams develop and small teams disrupt science and technology

Lingfei Wu[1,2], Dashun Wang[3,4,5] & James A. Evans[1,2,6]*

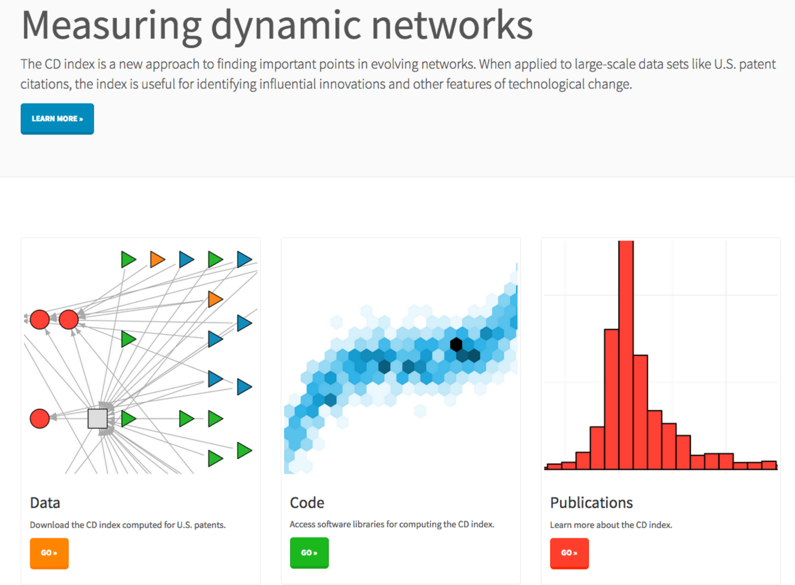
One of the most universal trends in science and technology today is the growth of large teams in all areas, as solitary researchers and small teams diminish in prevalence[1–3]. Increases in team size have been attributed to the specialization of scientific activities[3], improvements in communication technology[4,5], or the complexity of modern problems that require interdisciplinary solutions[6–8]. This shift in team size raises the question of whether and how the character of the science and technology produced by large teams differs from that of small teams. Here we analyse more than 65 million papers, patents and software products that span the period 1954–2014, and demonstrate that across this period smaller teams have tended to disrupt science and technology with new ideas and opportunities, whereas larger teams have tended to develop existing ones. Work from larger teams builds on more-recent and popular developments, and attention to their work comes immediately. By contrast, contributions by smaller teams search more deeply into the past, are viewed as disruptive to science and technology and succeed further into the future—if at all. Observed differences between small and large teams are magnified for higher-impact work, with small teams known for disruptive work and large teams for developing work. Differences in topic and research design account for a small part of the relationship between team size and disruption; most of the effect occurs at the level of the individual, as people move between smaller and larger teams. These results demonstrate that both small and large teams are essential to a flourishing ecology of science and technology, and suggest that, to achieve this, science policies should aim to support a diversity of

difference between two well-known articles: one about self-organized criticality[17] (the BTW model, after the authors' initials) and another about Bose–Einstein condensation[18] (for which Wolfgang Ketterle was awarded the 2001 Nobel Prize in Physics) (Fig. 1, Extended Data Fig. 1b). The two articles have received a similar number of citations, but most research subsequent to the BTW-model article has cited only the model itself without mentioning references from the article. By contrast, the Bose–Einstein condensation article is almost always co-cited with Bose[19], Einstein[20] and other antecedents. The difference between the two papers is reflected not in citation counts but in whether they suggested or solved scientific problems—whether they disrupted or developed existing scientific ideas, respectively[21]. The BTW model launched new streams of research, whereas the experimental realization of Bose–Einstein condensation elaborated upon possibilities that had previously been posed.

To systematically evaluate the role that small and large teams have in unfolding scientific and technological advances, we collected large-scale datasets from three related but distinct domains (see Methods): (1) the Web of Science (WOS) database that contains more than 42 million articles published between 1954 and 2014, and 611 million citations among them; (2) 5 million patents granted by the US Patent and Trademark Office from 1976 to 2014, and 65 million citations added by patent applicants; (3) 16 million software projects and 9 million forks to them on GitHub (2011–2014), a popular web platform that allows users to collaborate on the same code repository and 'cite' other repositories by copying and building on their code.

For each dataset, we assess the degree to which each work disrupts

Website



Measuring dynamic networks

The CD index is a new approach to finding important points in evolving networks. When applied to large-scale data sets like U.S. patent citations, the index is useful for identifying influential innovations and other features of technological change.

Data

Download the CD index computed for U.S. patents.

Code

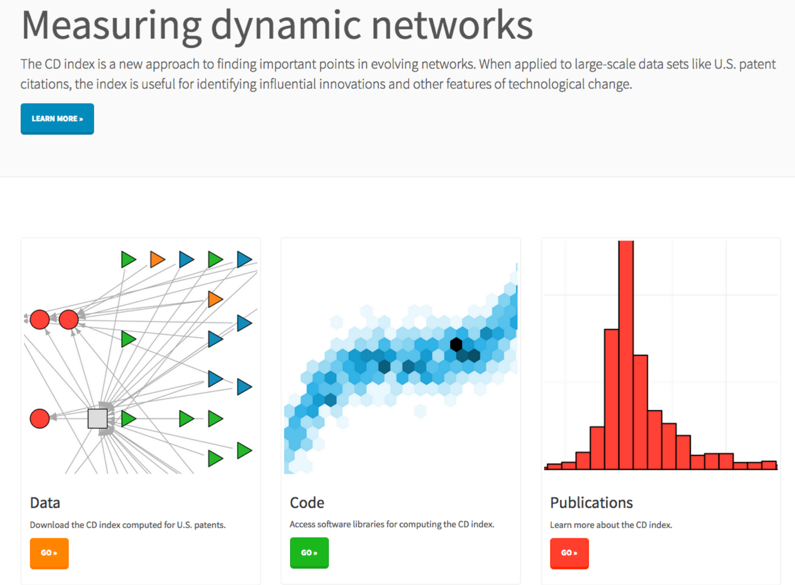
Access software libraries for computing the CD index.

Publications

Learn more about the CD index.

Appendix

Website



Measuring dynamic networks

The CD index is a new approach to finding important points in evolving networks. When applied to large-scale data sets like U.S. patent citations, the index is useful for identifying influential innovations and other features of technological change.

Data

Download the CD index computed for U.S. patents.

Code

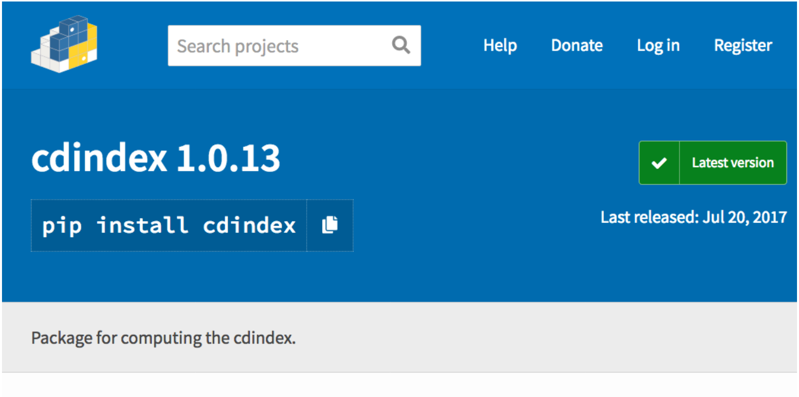
Access software libraries for computing the CD index.

Publications

Learn more about the CD index.

PyPI

```
pip install cdindex
```



cdindex 1.0.13

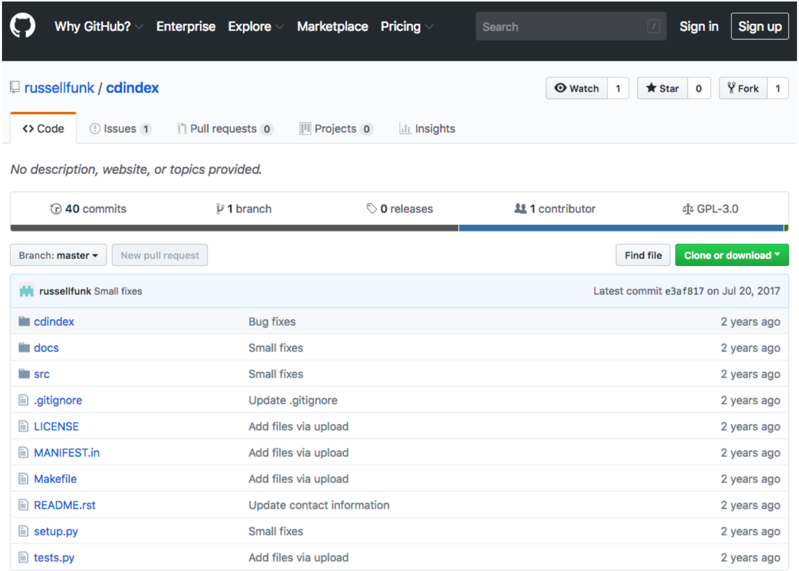
```
pip install cdindex
```

Last released: Jul 20, 2017

Package for computing the cdindex.

GitHub

https://github.com/russellfunk/cdindex



russellfunk / cdindex

No description, website, or topics provided.

40 commits | 1 branch | 0 releases | 1 contributor | GPL-3.0

russellfunk Small fixes — Latest commit e3af817 on Jul 20, 2017

cdindex	Bug fixes	2 years ago
docs	Small fixes	2 years ago
src	Small fixes	2 years ago
.gitignore	Update .gitignore	2 years ago
LICENSE	Add files via upload	2 years ago
MANIFEST.in	Add files via upload	2 years ago
Makefile	Add files via upload	2 years ago
README.rst	Update contact information	2 years ago
setup.py	Small fixes	2 years ago
tests.py	Add files via upload	2 years ago