A Weighted Correlation Index for Rankings with Ties

Sebastiano Vigna
Università degli Studi di Milano

The problem

- Understanding the correlation between different rankings
- Why do correlation measurements between centrality measures are so flaky?
- Taking care of ties is essential (indegree)
- Rank differences between important elements *should be more relevant*
- Large-scale target (whole graphs), not small sets of results

As a Motivation

- Indegree
- Bavelas's Closeness
- Harmonic Centrality
- PageRank
- Katz

Geometric Centralities

- Bavelas's closeness (1948): $\frac{1}{\sum_y d(y,x)}$

- Harmonic centrality (1965): $\sum_{y \neq x} \frac{1}{d(y,x)}$

- We're just moving from a denormalized, reciprocated arithmetic mean to a denormalized, reciprocated harmonic mean

Spectral Centralities

- Katz (1951): $\mathbf{1} \sum_{k \geq 0} \alpha^k G^k$
- PageRank: $\mathbf{1}/n \sum_{k \geq 0} \alpha^k \bar{G}^k$

Indegree	PageRank	Katz	Harmonic	Closeness
United States	United States	United States	United States	**Kharqan Rural District**
List of sovereign states	Animal	List of sovereign states	United Kingdom	**Talageh-ye Sofla**
Animal	List of sovereign states	United Kingdom	World War II	**Talageh-ye Olya**
England	France	France	France	**Greatest Remix Hits (Whigfield album)**
France	Germany	Animal	Germany	**Suzhou HSR New Town**
Association football	Association football	World War II	Association football	**Suzhou Lakeside New City**
United Kingdom	England	England	English language	**Mepirodipine**
Germany	India	Association football	China	**List of MPs ... M–N**
Canada	United Kingdom	Germany	Canada	**List of MPs ... O–R**
World War II	Canada	Canada	India	**List of MPs ... S–T**
India	Arthropod	India	**Latin**	**List of MPs ... U–Z**
Australia	Insect	Australia	World War I	**List of MPs ... J–L**
London	World War II	London	England	**List of MPs ... C**
Japan	Japan	Italy	Italy	**List of MPs ... F–I**
Italy	Australia	Japan	**Russia**	**List of MPs ... A–B**
Arthropod	Village	New York City	**Europe**	**List of MPs ... D–E**
Insect	Italy	English language	Australia	**Esmaili-ye Sofla**
New York City	Poland	China	**European Union**	**Esmaili-ye Olya**
English language	English language	Poland	**Catholic Church**	**Levels of organization (ecology)**
Village	**Nationa Reg. of Hist. Places**	World War I	London	**Jacques Moeschal (architect)**

Table 1: Top 20 pages of the English version of Wikipedia following five different centrality measures.

	Ind.	PR	Katz	Harm.	Cl.
Indegree	1	0.75	0.90	0.62	0.55
PageRank	0.75	1	0.75	0.61	0.56
Katz	0.90	0.75	1	0.70	0.62
Harmonic	0.62	0.61	0.70	1	0.92
Closeness	0.55	0.56	0.62	0.92	1

	Ind.	PR	Katz	Harm.	Cl.
Indegree	1	0.31	0.63	0.24	0.06
PageRank	0.31	1	0.27	0.10	0.10
Katz	0.63	0.27	1	0.50	0.20
Harmonic	0.24	0.10	0.50	1	0.65
Closeness	0.06	0.10	0.20	0.65	1

Indegree	PageRank	Katz	Harmonic	Closeness
Carl Linnaeus	Carl Linnaeus	Carl Linnaeus	Aristotle	**Noël Bernard (botanist)**
Aristotle	Aristotle	Aristotle	Albert Einstein	**Charles Coquelin**
Thomas Jefferson	Thomas Jefferson	Thomas Jefferson	Thomas Jefferson	**Markku Kivinen**
Margaret Thatcher	Charles Darwin	Albert Einstein	Charles Darwin	**Angiolo Maria Colomboni**
Plato	Plato	Charles Darwin	Thomas Edison	**Om Prakash (historian)**
Charles Darwin	Albert Einstein	Karl Marx	**Alexander Graham Bell**	**Michel Mandjes**
Karl Marx	Karl Marx	Plato	**Nikola Tesla**	**Kees Posthumus**
Albert Einstein	Pliny the Elder	Margaret Thatcher	**William James**	**F. Wolfgang Schnell**
Vladimir Lenin	Vladimir Lenin	Vladimir Lenin	Isaac Newton	**Christof Ebert**
Sigmund Freud	Johann Wolfgang von Goethe	Isaac Newton	Karl Marx	**Reese Prosser**
J. R. R. Tolkien	Margaret Thatcher	Ptolemy	**Charles Sanders Peirce**	**David Tulloch**
Johann Wolfgang von Goethe	Ptolemy	Johann Wolfgang von Goethe	Noam Chomsky	**Kim Hawtrey**
Spider-Man	Sigmund Freud	Pliny the Elder	**Enrico Fermi**	**Patrick J. Miller**
Pliny the Elder	Isaac Newton	Benjamin Franklin	Ptolemy	**Mikel King**
Benjamin Franklin	Benjamin Franklin	J. R. R. Tolkien	**John Dewey**	**Albert Perry Brigham**
Leonardo da Vinci	J. R. R. Tolkien	Thomas Edison	Johann Wolfgang von Goethe	**Gordon Wagner (economist)**
Isaac Newton	Immanuel Kant	Sigmund Freud	**Bertrand Russell**	**George Henry Chase**
Ptolemy	Leonardo da Vinci	Immanuel Kant	Plato	**Charles C. Horn**
Immanuel Kant	**Pierre André Latreille**	Leonardo da Vinci	**John von Neumann**	**Paul Goldstene**
George Bernard Shaw	Thomas Edison	Noam Chomsky	Vladimir Lenin	**Robert Stanton Avery**

Indegree	PageRank	Katz	Harmonic	Closeness
Martini (cocktail)	Martini (cocktail)	Irish coffee	Irish coffee	**Magie Noir**
Piña colada	Caipirinha	Caipirinha	Caipirinha	**Batini (drink)**
Mojito	Mojito	Martini (cocktail)	Kir (cocktail)	**Scorpion bowl**
Caipirinha	Piña colada	Piña colada	Martini (cocktail)	**Poinsettia (cocktail)**
Cuba Libre	Irish coffee	Kir (cocktail)	Piña colada	Irish coffee
Irish coffee	Kir (cocktail)	Mojito	Mojito	Caipirinha
Singapore Sling	Cosmopolitan (cocktail)	Mai Tai	Beer cocktail	Kir (cocktail)
Manhattan (cocktail)	Manhattan (cocktail)	Cuba Libre	Shaken, not stirred	Martini (cocktail)
Windle (sidecar)	IBA Official Cocktail	Singapore Sling	Pisco Sour	Piña colada
Cosmopolitan (cocktail)	Beer cocktail	Long Island Iced Tea	Mai Tai	Mojito
Mai Tai	Mai Tai	Shaken, not stirred	Spritz (alcoholic beverage)	Beer cocktail
IBA Official Cocktail	Singapore Sling	Beer cocktail	Long Island Iced Tea	Shaken, not stirred
Kir (cocktail)	Cuba Libre	Manhattan (cocktail)	Sazerac	Mai Tai
Shaken, not stirred	**Tom Collins**	Cosmopolitan (cocktail)	Fizz (cocktail)	Spritz (alcoholic beverage)
Beer cocktail	Long Island Iced Tea	Windle (sidecar)	Flaming beverage	Pisco Sour
Pisco Sour	Sour (cocktail)	Pisco Sour	Cuba Libre	Long Island Iced Tea
Long Island Iced Tea	Shaken, not stirred	White Russian (cocktail)	Wine cocktail	Sazerac
Sour (cocktail)	**Negroni**	IBA Official Cocktail	Singapore Sling	Flaming beverage
White Russian (cocktail)	Flaming beverage	Moscow mule	Moscow mule	Fizz (cocktail)
Vesper (cocktail)	**Lillet**	Vesper (cocktail)	White Russian (cocktail)	Wine cocktail

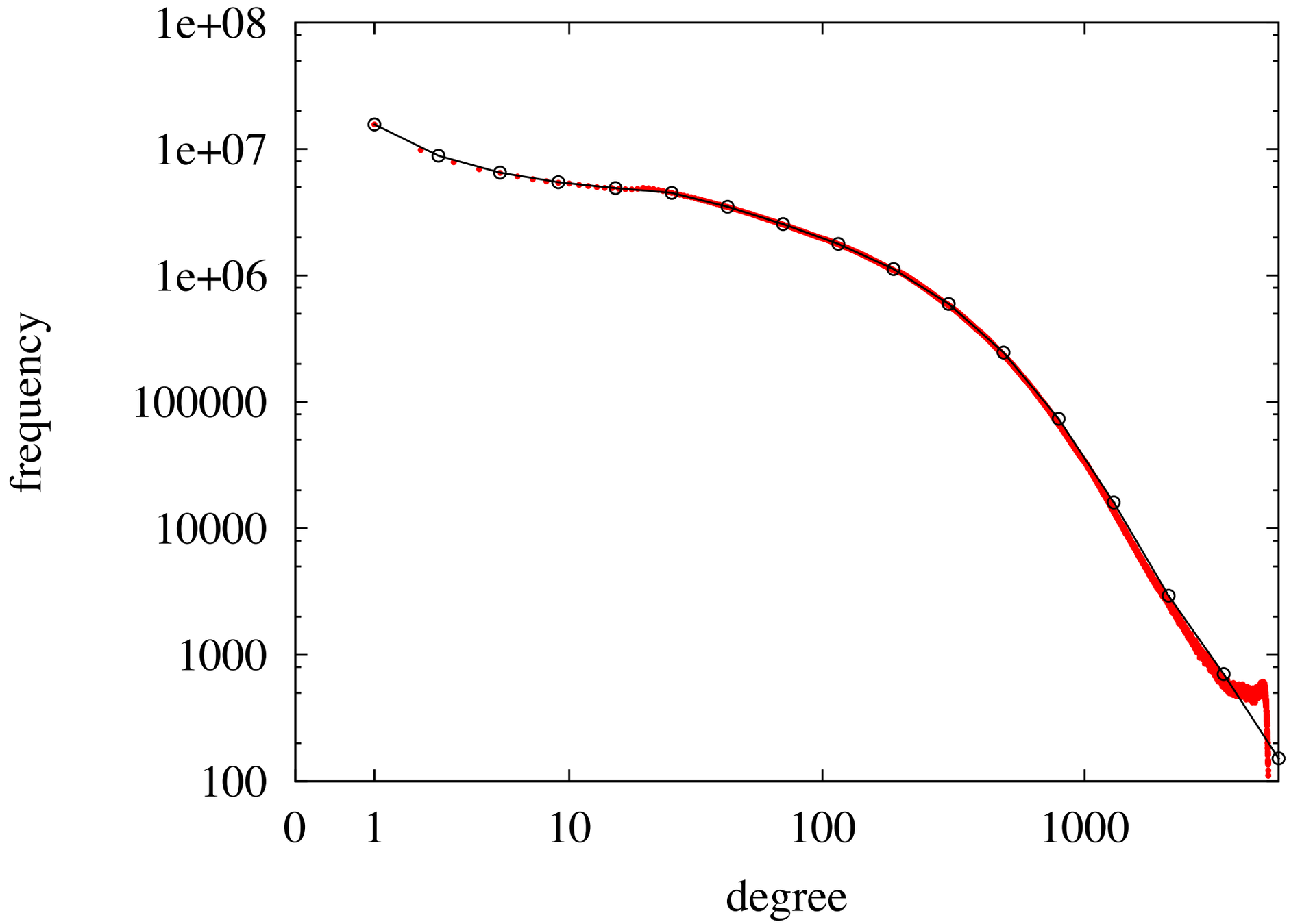
Kendall's τ 1938

- Scores ***r***, ***s*** (distinct)
- Concordances: pairs (i, j), $i < j$, such that the scores for i and j in ***r*** and ***s*** are in the same order
- τ: Concordances minus discordances divided by concordances plus discordances (i.e., the number of ordered pairs)
- Note: if you skip ties on both sides you get Goodman–Kruskal's γ

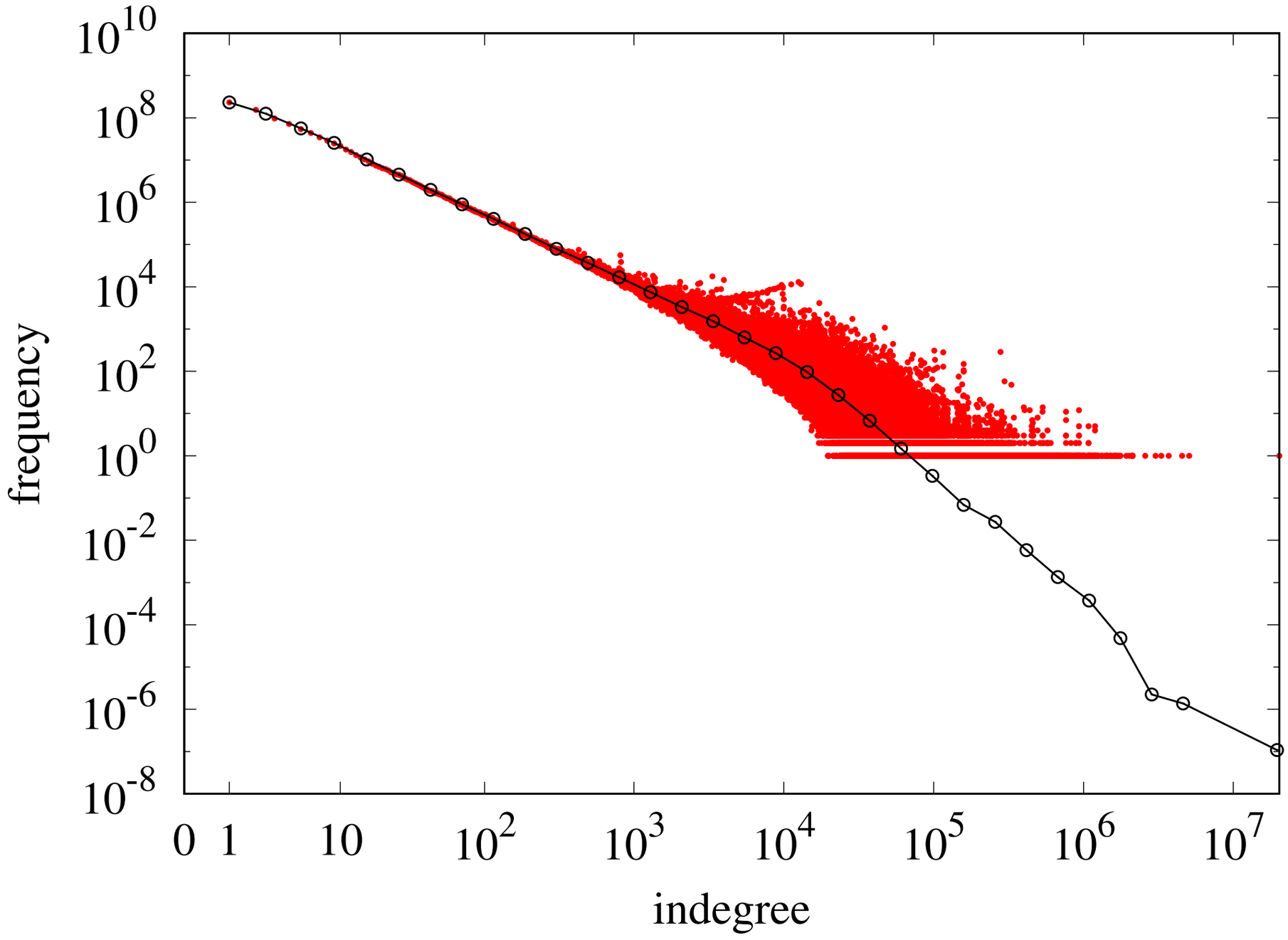
Ties Are Important

- >99% nodes in a typical social/web graph are in a degree tie
- Ties cannot be solved by random assignment
- <0,0,0,...,1,1,1,...> and <1,1,1,...,2,2,2,...> give correlation ≈0.5
- When you compare network rankings, you're almost always working in the 0.7-to-1 region

Facebook (2011)



.eu (2015, 10^9 pages)



Kendall's τ 1945

- Starts from Daniels (1944): every correlation is a cosine similarity

$$\langle \boldsymbol{r}, \boldsymbol{s} \rangle := \sum_{i<j} \operatorname{sgn}(r_i - r_j) \operatorname{sgn}(s_i - s_j)$$

- "Norm"

$$\|\boldsymbol{r}\| := \sqrt{\langle \boldsymbol{r}, \boldsymbol{r} \rangle}$$

- Cosine similarity!

$$\tau(\boldsymbol{r}, \boldsymbol{s}) := \frac{\langle \boldsymbol{r}, \boldsymbol{s} \rangle}{\|\boldsymbol{r}\| \cdot \|\boldsymbol{s}\|}$$

Now

$$\langle \boldsymbol{r}, \boldsymbol{s} \rangle_w := \sum_{i<j} \operatorname{sgn}(r_i - r_j) \operatorname{sgn}(s_i - s_j) w(i, j)$$

$$|\langle \boldsymbol{r}, \boldsymbol{s} \rangle_w| \leq \|\boldsymbol{r}\|_w \|\boldsymbol{s}\|_w$$

$$\tau_w(\boldsymbol{r}, \boldsymbol{s}) := \frac{\langle \boldsymbol{r}, \boldsymbol{s} \rangle_w}{\|\boldsymbol{r}\|_w \cdot \|\boldsymbol{s}\|_w}$$

Related

- Shieh [St. & Pr. L. 1998]: weight the pair (i, j) with w_{ij}
- No ties!
- Yilmaz, Aslam & Robertson [SIGIR 2008]: $w_{ij} := 1/j$ (motivated by probability)
- Kumar & Vassilvitskii [WWW 2010]: w_{ij} depending on similarity, ties broken randomly
- Iman & Conover [Technometrics 1987]: Spearman's correlation between *Savage scores* $H_n - H_{i-1}$
- Farnoud [2012]: weight adjacent transpositions $(i\ i+1)$ and compute minimum weight

Decoupling Rank and Weight

$$\langle \boldsymbol{r}, \boldsymbol{s} \rangle_{\rho,w} := \sum_{i<j} \operatorname{sgn}(r_i - r_j) \operatorname{sgn}(s_i - s_j) w(\rho(i), \rho(j))$$

$$\tau_{\bullet,w}(\boldsymbol{r}, \boldsymbol{s}) := \frac{\tau_{\rho_{\boldsymbol{r},\boldsymbol{s}},w}(\boldsymbol{r}, \boldsymbol{s}) + \tau_{\rho_{\boldsymbol{s},\boldsymbol{r}},w}(\boldsymbol{r}, \boldsymbol{s})}{2}$$

- $\rho_{\boldsymbol{r},\boldsymbol{s}}$ is the ranking induced by $\boldsymbol{r}$ and $\boldsymbol{s}$ in lexicographical order
- viceversa for $\rho_{\boldsymbol{s},\boldsymbol{r}}$

Computable Quickly

- $O(n \log n)$ variant of Knight's algorithm (highly parallelizable, distributable—it's a MergeSort)
- Works for any scheme $w(i, j) := f(i) \odot g(j)$ with suitable operation $\odot$ (e.g., addition, multiplication)
- We suggest *additive hyperbolic* weighting, weighting (i, j) by $1/(i+1)+1/(j+1)$: τ_h

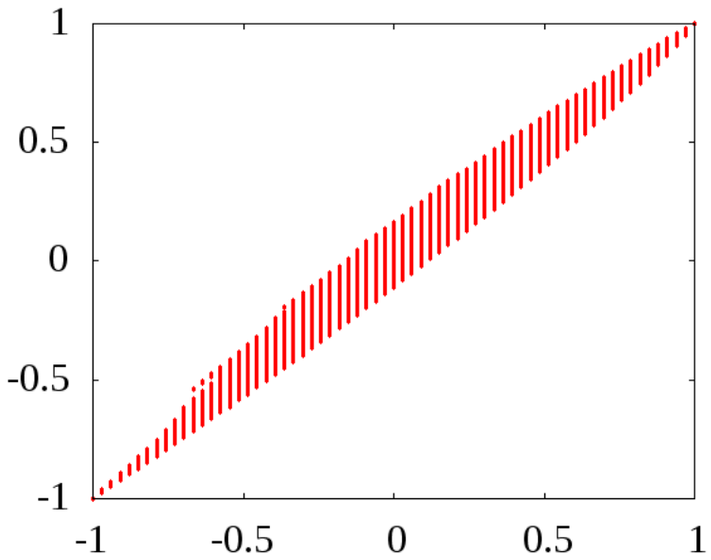
Correlation with Kendall's τ

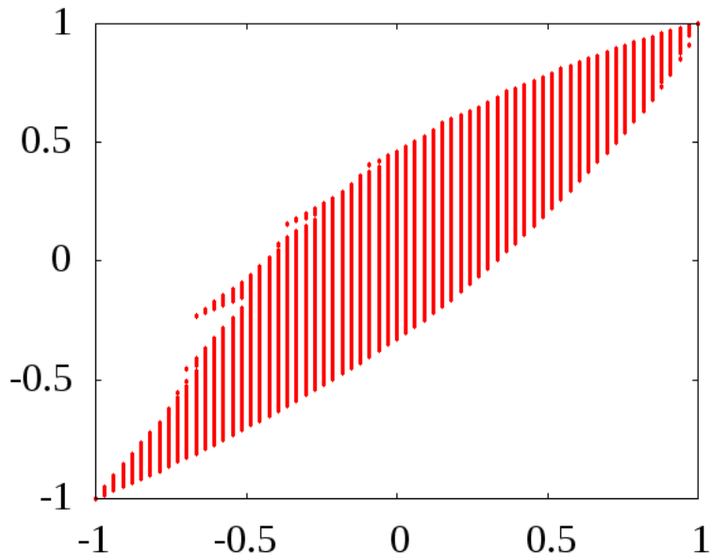
Logarithmic

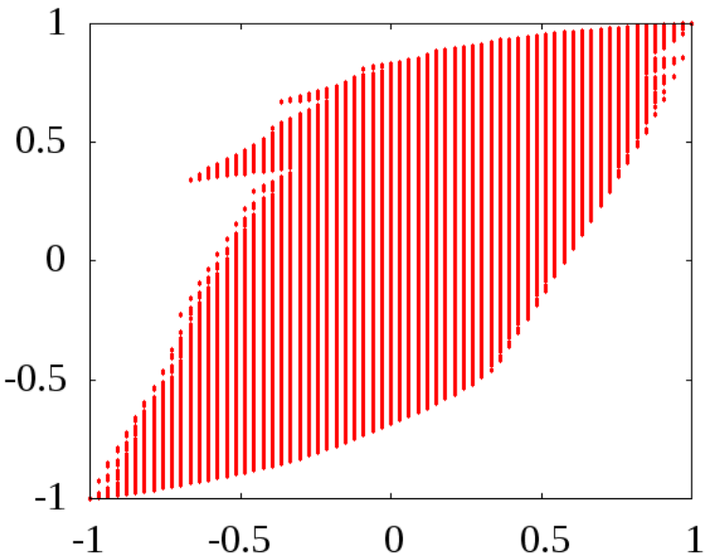
Hyperbolic

Quadratic

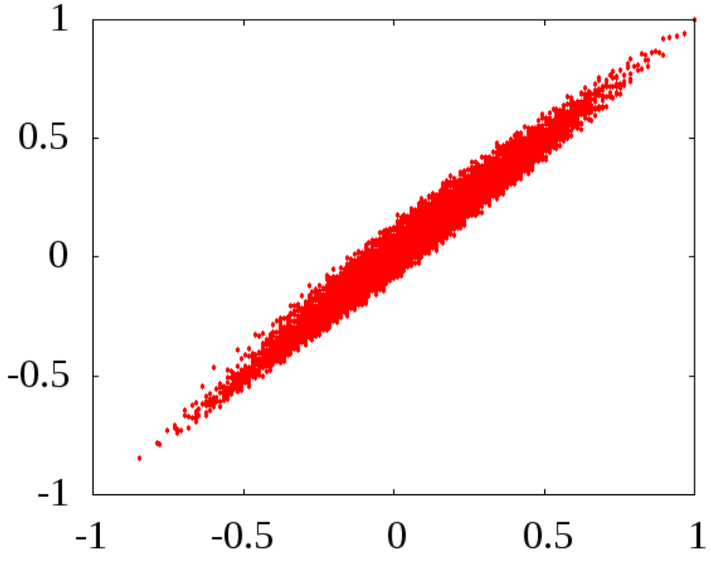
No ties

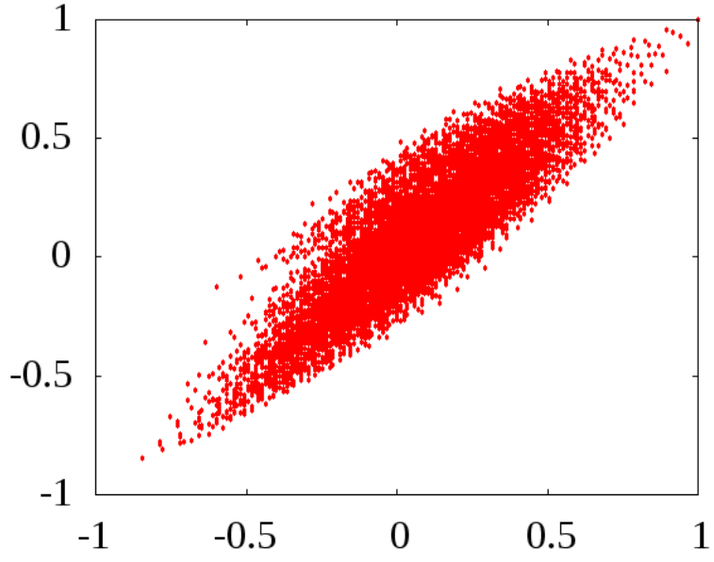


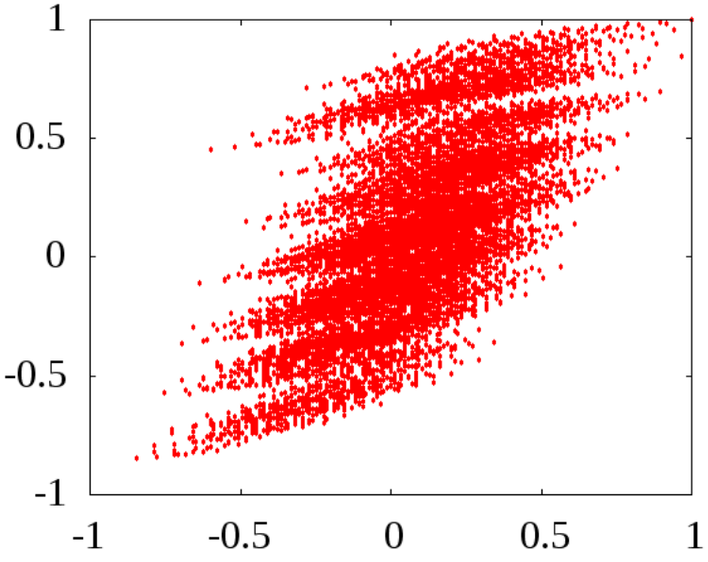




Ties







Wikipedia

τ

	Ind.	PR	Katz	Harm.	Cl.
Indegree	1	0.75	0.90	0.62	0.55
PageRank	0.75	1	0.75	0.61	0.56
Katz	0.90	0.75	1	0.70	0.62
Harmonic	0.62	0.61	0.70	1	0.92
Closeness	0.55	0.56	0.62	0.92	1

τ_h

	Ind.	PR	Katz	Harm.	Cl.
Indegree	1	0.95	0.98	0.90	0.27
PageRank	0.95	1	0.96	0.92	0.65
Katz	0.98	0.96	1	0.93	0.26
Harmonic	0.90	0.92	0.93	1	0.28
Closeness	0.27	0.65	0.26	0.28	1

Table 6: τ_h on Wikipedia.

Hollywood co-starship

Indegree	PageRank	Katz	Harmonic	Closeness
Shatner, William	Jeremy, Ron	Shatner, William	Sheen, Martin	**Östlund, Claes Göran**
Flowers, Bess	Hitler, Adolf	Sheen, Martin	Clooney, George	**Östlund, Catarina**
Sheen, Martin	**Kaufman, Lloyd**	Hanks, Tom	Jackson, Samuel L.	**von Preußen, Oskar Prinz**
Reagan, Ronald (I)	Bush, George W.	Williams, Robin (I)	Hopper, Dennis	**von Preußen, Georg Friedrich**
Clooney, George	Reagan, Ronald (I)	Clooney, George	Hanks, Tom	**von Mannstein, Robert Grund**
Jackson, Samuel L.	Clinton, Bill (I)	Reagan, Ronald (I)	Stone, Sharon (I)	**von Mannstein, Concha**
Williams, Robin (I)	Sheen, Martin	Willis, Bruce	Brosnan, Pierce	**von der Busken, Mart**
Hanks, Tom	**Rochon, Debbie**	Jackson, Samuel L.	Hitler, Adolf	**van der Putten, Thea**
Jeremy, Ron	**Kennedy, John F.**	Stone, Sharon (I)	**McDowell, Malcolm**	**de la Bruheze, Joel Albert**
Hitler, Adolf	Hopper, Dennis	Freeman, Morgan (I)	Williams, Robin (I)	**de la Bruheze, Emile**
Willis, Bruce	**Nixon, Richard**	Flowers, Bess	**De Niro, Robert**	**te Riele, Marloes**
Clinton, Bill (I)	**Estevez, Joe**	Brosnan, Pierce	Willis, Bruce	**de Reijer, Eric**
Freeman, Morgan (I)	Shatner, William	Douglas, Michael (I)	**Hopkins, Anthony**	**des Bouvrie, Jan**
Hopper, Dennis	Jackson, Samuel L.	Madonna (I)	Madonna (I)	**de Klijn, Judith**
Stone, Sharon (I)	**Stewart, Jon (I)**	Travolta, John	**Lee, Christopher (I)**	**de Freitas, Luís (II)**
Madonna (I)	**Carradine, David (I)**	Hopper, Dennis	Douglas, Michael (I)	**de Freitas, Luís (I)**
Bush, George W.	Asner, Edward	Ford, Harrison (I)	**Sutherland, Donald (I)**	**Zuu, Winnie Otondi**
Harris, Sam (II)	**Zirnkilton, Steven**	Asner, Edward	Freeman, Morgan (I)	**Zuu, Emmanuel Dahngbay**
Brosnan, Pierce	**Colbert, Stephen**	**MacLaine, Shirley**	**Stallone, Sylvester**	**Zilbersmith, Carla**
Travolta, John	**Madsen, Michael (I)**	Clinton, Bill (I)	Ford, Harrison (I)	**Zilber, Mac**

Table 10: Top 20 pages of the Hollywood co-starship graph.

Hollywood co-starship

τ

	Ind.	PR	Katz	Harm.	Cl.
Indegree	1	0.42	0.93	0.55	0.43
PageRank	0.42	1	0.36	0.10	0.18
Katz	0.93	0.36	1	0.61	0.49
Harmonic	0.55	0.10	0.61	1	0.86
Closeness	0.43	0.18	0.49	0.86	1

τ_h

	Ind.	PR	Katz	Harm.	Cl.
Indegree	1	0.90	0.98	0.91	0.10
PageRank	0.90	1	0.88	0.81	0.64
Katz	0.98	0.88	1	0.92	0.11
Harmonic	0.91	0.81	0.92	1	0.18
Closeness	0.10	0.64	0.11	0.18	1

Common Crawl Hosts

Indegree	PageRank	Katz	Harmonic	Closeness
wordpress.com	gmpg.org	wordpress.com	youtube.com	**0-p.com**
youtube.com	wordpress.com	youtube.com	en.wikipedia.org	**0-0-0-0-0-0-0.indahiphop.ru**
gmpg.org	youtube.com	gmpg.org	twitter.com	**0-0-1.i.tiexue.net**
en.wikipedia.org	**livejournal.com**	en.wikipedia.org	google.com	**0-00cigarettes.info**
tumblr.com	tumblr.com	tumblr.com	wordpress.org	**0-0mos00.hi5.com**
twitter.com	en.wikipedia.org	twitter.com	flickr.com	**0-0new0-0.hi5.com**
google.com	twitter.com	google.com	facebook.com	**0-0sunny0-0.hi5.com**
flickr.com	**networkadvertising.org**	flickr.com	**apple.com**	**0-1.i.tiexue.net**
rtalabel.org	**promodj.com**	rtalabel.org	vimeo.com	**0-1.sxsy.co**
wordpress.com	**skriptmail.de**	wordpress.com	creativecommons.org	**0-2.paparazziwannabe.com**
mp3shake.com	**parallels.com**	mp3shake.com	**amazon.com**	**0-311.cn**
w3schools.com	**tistory.com**	w3schools.com	**adobe.com**	**0-360.rukazan.ru**
domains.lycos.com	google.com	creativecommons.org	**myspace.com**	**0-5days.com**
staff.tumblr.com	miibeian.gov.cn	staff.tumblr.com	**w3.org**	**0-5days.net**
club.tripod.com	phpbb.com	domains.lycos.com	**bbc.co.uk**	**0-5kalibr.pdj.ru**
creativecommons.org	**blog.fc2.com**	club.tripod.com	**nytimes.com**	**0-9-0-4-4-9.promoradio.ru**
vimeo.com	**tw.yahoo.com**	vimeo.com	**yahoo.com**	**0-9-0-9.dbass.ru**
miibeian.gov.cn	w3schools.com	miibeian.gov.cn	**microsoft.com**	**0-9-0-9.promodj.ru**
facebook.com	wordpress.com	facebook.com	**guardian.co.uk**	**0-9-1125.i.tiexue.net**
phpbb.com	domains.lycos.com	phpbb.com	**imdb.com**	**0-9-7-16.software.informer.com**

Common Crawl Hosts

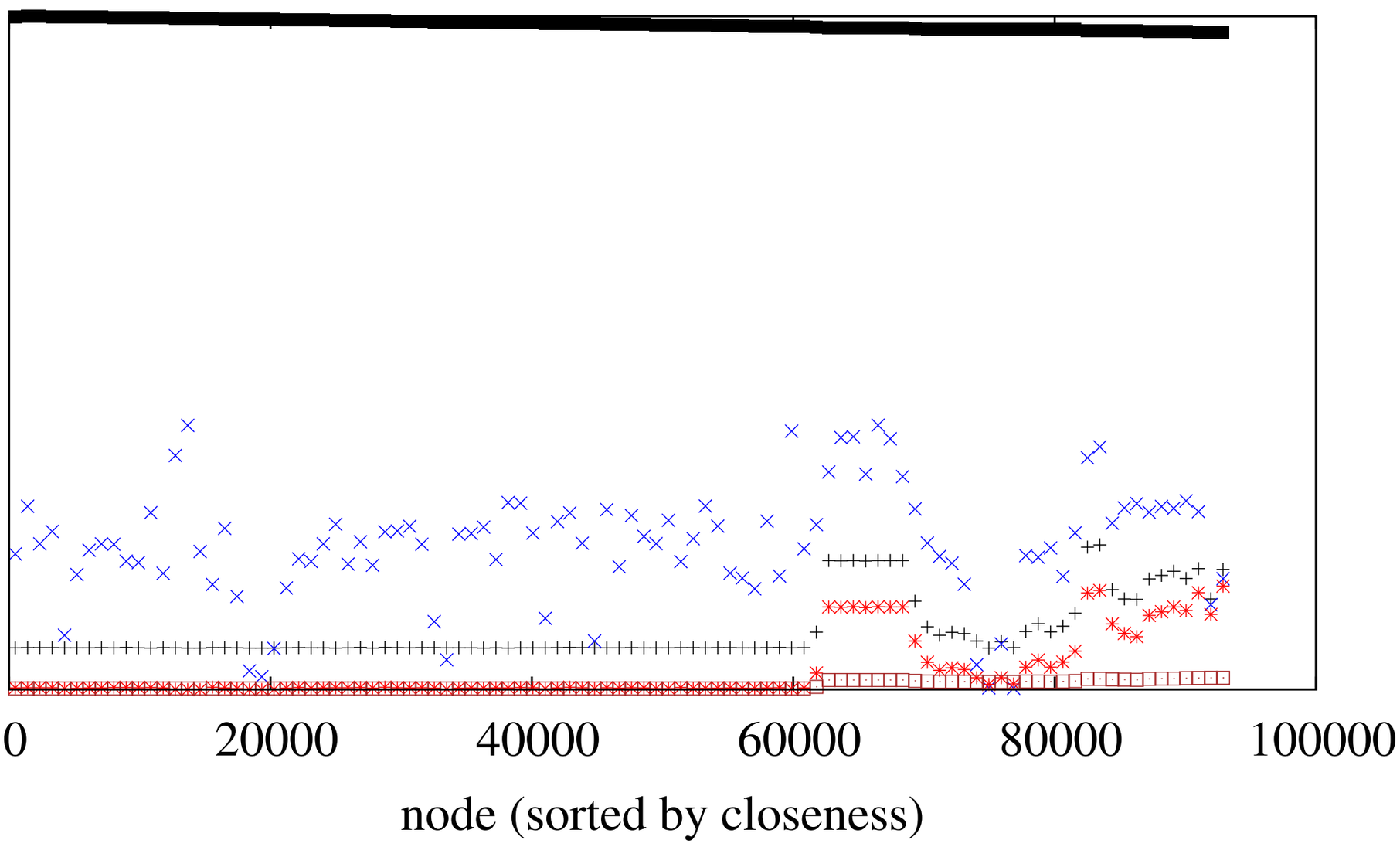
τ

	Ind.	PR	Katz	Harm.	Cl.
Indegree	1	0.71	0.89	0.61	0.54
PageRank	0.71	1	0.66	0.50	0.50
Katz	0.89	0.66	1	0.69	0.59
Harmonic	0.61	0.50	0.69	1	0.86
Closeness	0.54	0.50	0.59	0.86	1

τ_h

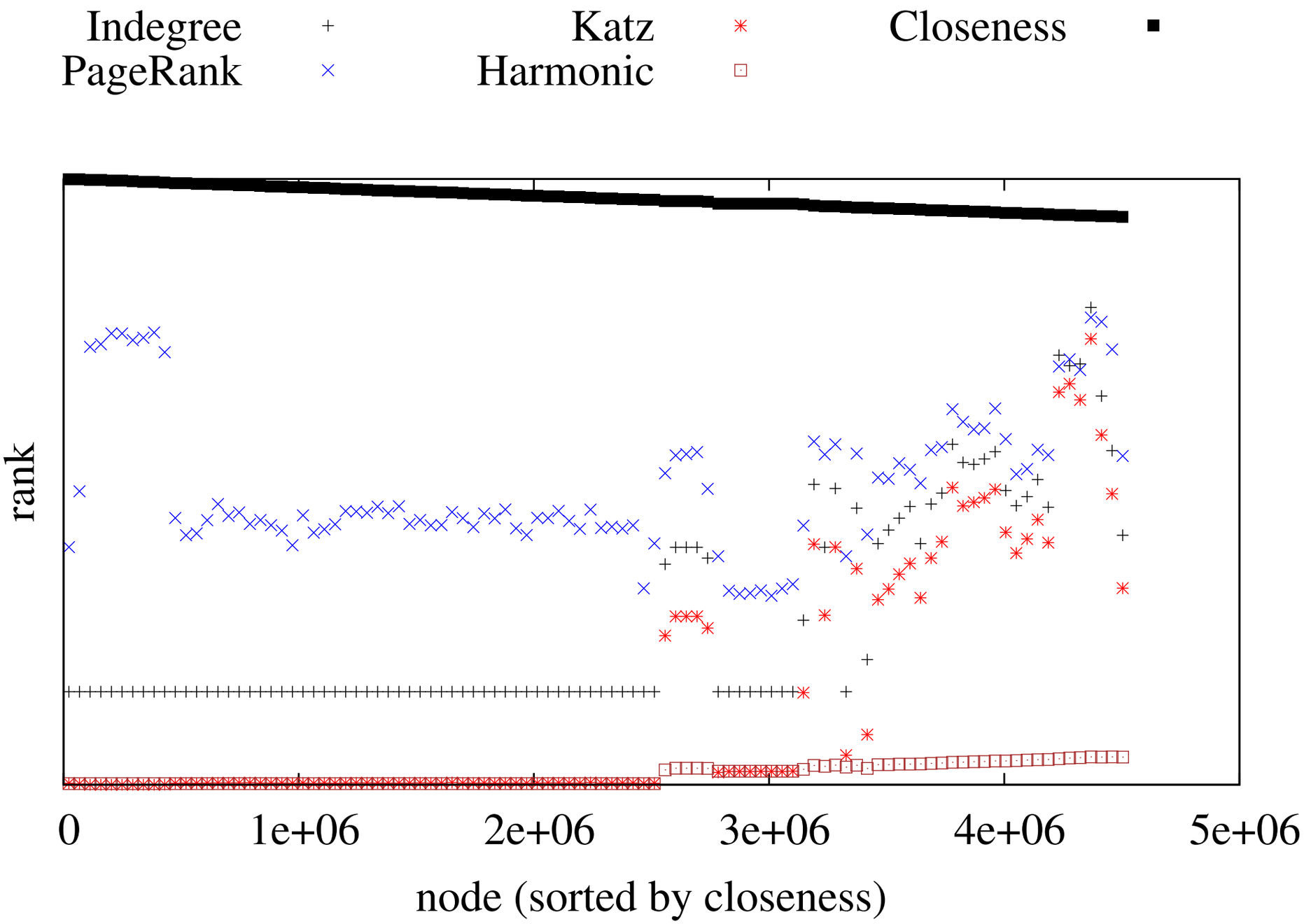
	Ind.	PR	Katz	Harm.	Cl.
Indegree	1	0.91	0.96	0.72	0.20
PageRank	0.91	1	0.90	0.81	0.69
Katz	0.96	0.90	1	0.78	0.15
Harmonic	0.72	0.81	0.78	1	0.35
Closeness	0.20	0.69	0.15	0.35	1

Why does closeness τ_h -correlate with PageRank?



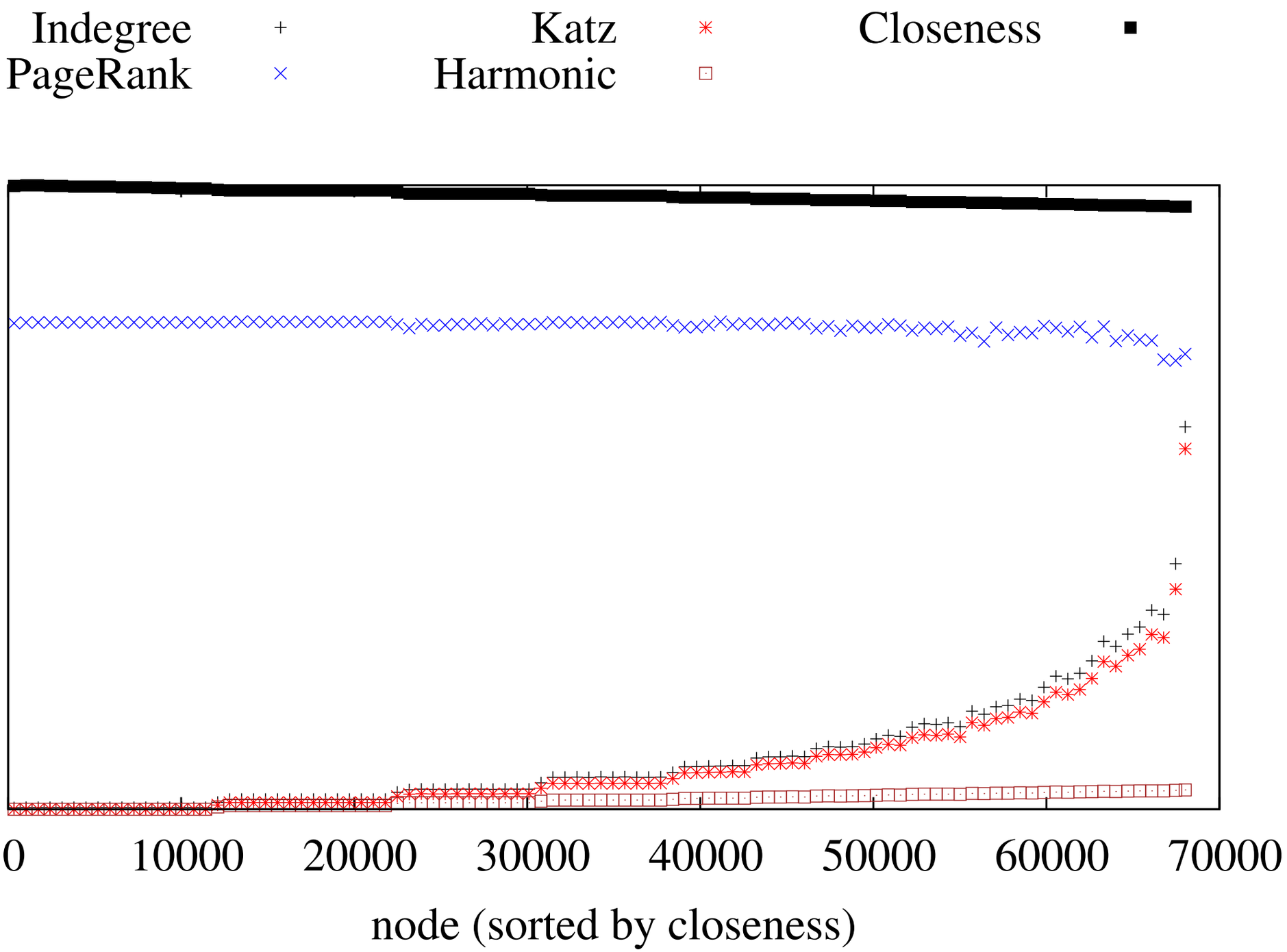
Rank of nodes unreachable from the giant component of Wikipedia

Why does closeness τ_h -correlate with PageRank?



Rank of nodes unreachable from the giant component of Common Crawl

Why does closeness τ_h -correlate with PageRank?



Rank of nodes unreachable from the giant component of Hollywood

Conclusions

- We believe τ_h is a new and valuable tool to understand rankings
- In general, the machinery around τ_w can be used to try easily new application-dependent weighted correlation indices on a large scale
- There are obvious elements of arbitrariness, but hyperbolic weighting is at the convergence of several previous proposals
- Implemented as `stats.weightedtau` in SciPy (easy to use!)
- Software, as usual, at `http://law.di.unimi.it/`