Session III

Case studies

Case-study: Biographies on Wikipedia

Aragon, P., Laniado, D., Kaltenbrunner, A., & Volkovich, Y. (2012, August). Biographical social networks on Wikipedia: a cross-cultural study of links that made history. In Proceedings of the eighth annual international symposium on Wikis and open collaboration (pp. 1-4).

Motivation

Is history made by great man and women or vice-versa?

- Unclear, but undoubtedly social connections shape history.

Wikipedia as global collective memory place...

- allows to extract from biographies how social links are recorded across cultures...
- to generate networks of links between biographical articles.

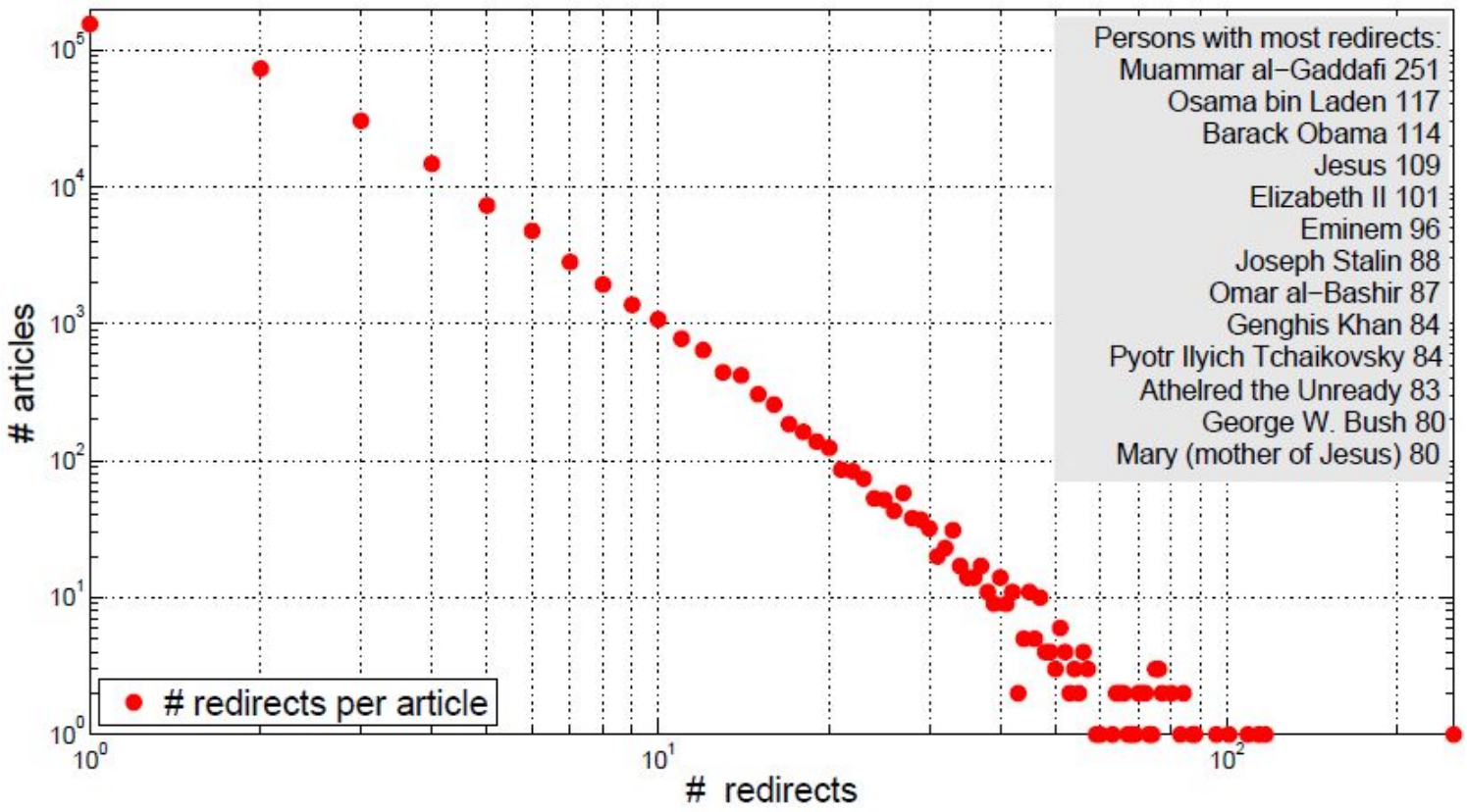
Research questions

- Who are the most central characters in these networks?
- Do culture related peculiarities exist?
- Which cultures are more similar?
- What is the shared knowledge about connections between persons across cultures?

Data extraction

- Selected the 15 largest language editions of Wikipedias
- Starting point: 296 511 biographies from the English Wikipedia (from DBpedia)
- Identified the corresponding articles (when existing) on the remaining 14 languages
- Generated a directed network for each language version:
  - nodes → persons
  - edges → links between the articles of the corresponding persons

Track redirects



Manage alternative titles of articles

Structural metrics

lang	N	K	$\langle C \rangle$	% GC	$\langle d \rangle$	r	d_{max}
en	198 190	928 339	0.03	95%	6.53	0.17	43
de	62 402	260 889	0.05	94%	6.83	0.14	33
fr	51 811	283 453	0.06	96%	6.11	0.15	36
it	35 756	190 867	0.06	95%	6.28	0.14	42
es	34 828	169 302	0.06	97%	6.29	0.16	36
ja	26 155	109 081	0.08	96%	6.47	0.20	26
nl	24 496	76 651	0.08	94%	7.91	0.18	37
pt	23 705	85 295	0.07	94%	6.98	0.18	45
sv	23 085	60 745	0.07	91%	8.27	0.20	46
pl	22 438	50 050	0.08	85%	8.94	0.16	43
fi	18 594	44 941	0.07	87%	7.80	0.17	30
no	18 423	49 303	0.09	83%	8.31	0.22	48
ru	16 403	34 436	0.06	87%	9.10	0.10	35
zh	11 715	44 739	0.17	91%	7.20	0.20	32
ca	11 027	42 321	0.09	93%	7.14	0.17	32

- $N, K \rightarrow$ number of (not isolated) nodes and edges
- $\langle C \rangle \rightarrow$ average clustering coefficient
- $GC \rightarrow$ percentage of nodes in the giant component
- $r \rightarrow$ reciprocity
- $\langle d \rangle \rightarrow$ average path-length between nodes
- $d_{max} \rightarrow$ maximal distance between two nodes in the network

Centrality measures

person	in-degree	out-degree		betw.	PageRank	
George W. Bush	2123	89	(107)	(1)	0.00209	(1)
Barack Obama	1677	51	(710)	(8)	0.00162	(2)
Bill Clinton	1660	74	(205)	(4)	0.00156	(4)
Ronald Reagan	1652	90	(103)	(2)	0.00156	(3)
Adolf Hitler	1407	119	(26)	(3)	0.00149	(5)
Richard Nixon	1299	86	(127)	(7)	0.00136	(6)
William Shakespeare	1229	25	(4203)	(63)	0.00113	(9)
John F. Kennedy	1208	104	(53)	(5)	0.00123	(8)
Franklin D. Roosevelt	1052	71	(237)	(15)	0.00131	(7)
Lyndon B. Johnson	1000	106	(50)	(12)	0.00108	(11)
Jimmy Carter	953	80	(158)	(9)	0.00113	(10)
Elvis Presley	948	82	(142)	(27)	0.00063	(24)
Pope John Paul II	941	59	(444)	(11)	0.00083	(18)
Dwight D. Eisenhower	891	55	(564)	(22)	0.00095	(14)
Frank Sinatra	882	108	(47)	(18)	0.00056	(28)
George H. W. Bush	878	87	(118)	(19)	0.00096	(13)
Abraham Lincoln	846	54	(593)	(40)	0.00089	(16)
Bob Dylan	835	151	(11)	(14)	0.00055	(30)
Winston Churchill	748	84	(136)	(10)	0.00092	(15)
Harry S. Truman	743	81	(145)	(24)	0.00099	(12)
Joseph Stalin	723	69	(265)	(43)	0.00089	(17)
Michael Jackson	663	71	(237)	(34)	0.00042	(51)
Elizabeth II	653	52	(665)	(6)	0.00074	(19)
Jesus	572	38	(1595)	(51)	0.00068	(20)
Hillary Rodham Clinton	554	87	(118)	(32)	0.00063	(25)

Centrality (English Wikipedia)

person	in-degree	out-degree		betw.	PageRank	
George W. Bush	2123	89	(107)	(1)	0.00209	(1)
Barack Obama	1677	51	(710)	(8)	0.00162	(2)
Bill Clinton	1660	74	(205)	(4)	0.00156	(4)
Ronald Reagan	1652	90	(103)	(2)	0.00156	(3)
Adolf Hitler	1407	119	(26)	(3)	0.00149	(5)
Richard Nixon	1299	86	(127)	(7)	0.00136	(6)
William Shakespeare	1229	25	(4203)	(63)	0.00113	(9)
John F. Kennedy	1208	104	(53)	(5)	0.00123	(8)
Franklin D. Roosevelt	1052	71	(237)	(15)	0.00131	(7)
Lyndon B. Johnson	1000	106	(50)	(12)	0.00108	(11)
Jimmy Carter	953	80	(158)	(9)	0.00113	(10)
Elvis Presley	948	82	(142)	(27)	0.00063	(24)
Pope John Paul II	941	59	(444)	(11)	0.00083	(18)
Dwight D. Eisenhower	891	55	(564)	(22)	0.00095	(14)
Frank Sinatra	882	108	(47)	(18)	0.00056	(28)
George H. W. Bush	878	87	(118)	(19)	0.00096	(13)
Abraham Lincoln	846	54	(593)	(40)	0.00089	(16)
Bob Dylan	835	151	(11)	(14)	0.00055	(30)
Winston Churchill	748	84	(136)	(10)	0.00092	(15)
Harry S. Truman	743	81	(145)	(24)	0.00099	(12)
Joseph Stalin	723	69	(265)	(43)	0.00089	(17)
Michael Jackson	663	71	(237)	(34)	0.00042	(51)
Elizabeth II	653	52	(665)	(6)	0.00074	(19)
Jesus	572	38	(1595)	(51)	0.00068	(20)
Hillary Rodham Clinton	554	87	(118)	(32)	0.00063	(25)

Centrality (all Wikipedia)

lang	#1	#2	#3	#4	#5
en	George W. Bush	Ronald Reagan	Adolf Hitler	Bill Clinton	John F. Kennedy
de	Adolf Hitler	George W. Bush	Martin Luther King, Jr	Barack Obama	Frank Sinatra
fr	Adolf Hitler	George W. Bush	William Shakespeare	Barack Obama	Jacques Chirac
it	Frank Sinatra	George W. Bush	Pope John Paul II	Michael Jackson	Elton John
es	Michael Jackson	Fidel Castro	William Shakespeare	Che Guevara	Adolf Hitler
ja	Adolf Hitler	Michael Jackson	Ronald Reagan	Yukio Mishima	Barack Obama
nl	Elvis Presley	Adolf Hitler	Bill Clinton	Joseph Stalin	William Shakespeare
pt	Michael Jackson	Richard Wagner	Adolf Hitler	Ronald Reagan	David Bowie
sv	George W. Bush	Winston Churchill	Elizabeth II	Michael Jackson	Adolf Hitler
pl	Elizabeth II	Pope John Paul II	Margaret Thatcher	George W. Bush	Ronald Reagan
fi	Barack Obama	Adolf Hitler	Michael Jackson	George W. Bush	Benito Mussolini
no	Marilyn Monroe	Adolf Hitler	John F. Kennedy	Bob Dylan	Bill Clinton
ru	William Shakespeare	Napoleon II	Kenneth Branagh	Elton John	Joseph Stalin
zh	Chiang Kai-Shek	William Shakespeare	Barack Obama	Deng Xiaoping	Adolf Hitler
ca	Adolf Hitler	Che Guevara	Juan Carlos I	Michael Schumacher	Juan Manuel Fangio

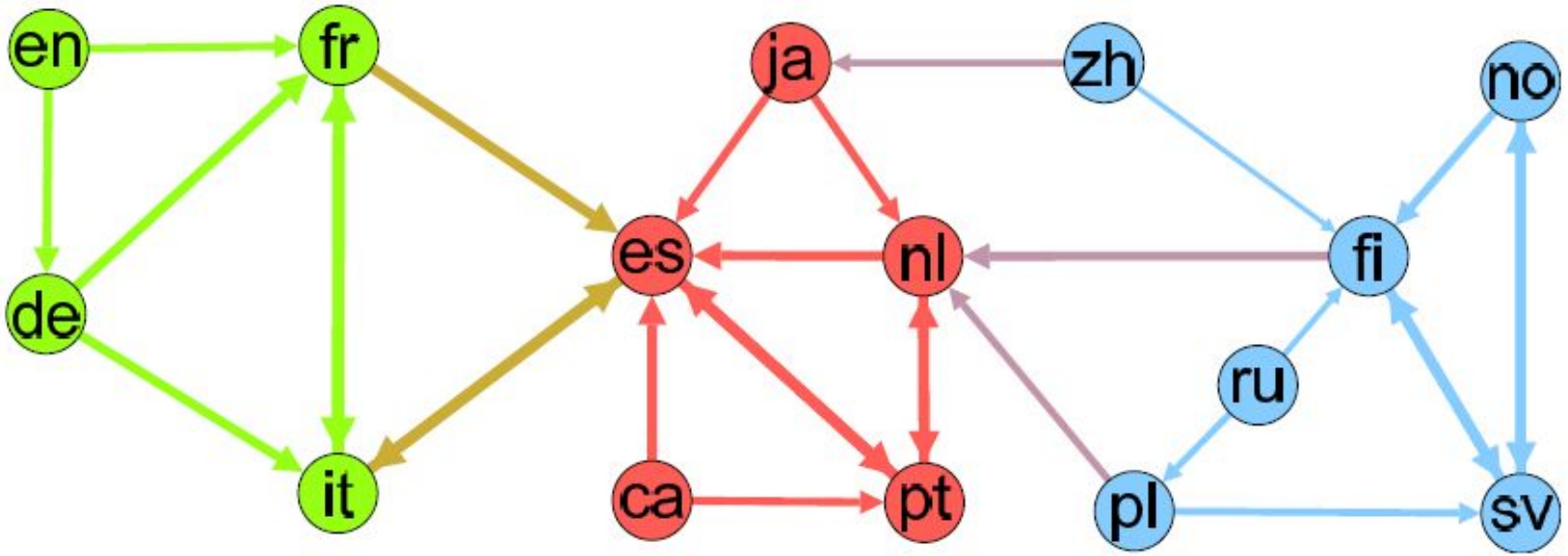
- We find political leaders, revolutionaries, famous musicians, writers and actors.
- Hitler, Bush, Obama dominate in almost all top rankings.
- Top ranked in many languages reflect country specificity.

Similarity among language editions

	ca	de	en	es	fi	fr	it	ja	nl	no	pl	pt	ru	sv	zh
ca	-	.05	.03	**.12**	.09	.07	.08	.07	.10	.08	.06	**.10**	.06	.09	.06
de	.05	-	.11	.11	.07	**.13**	**.12**	.08	.09	.06	.06	.08	.04	.08	.03
en	.03	**.11**	-	.09	.03	**.10**	.08	.05	.05	.03	.03	.05	.02	.04	.02
es	.12	.11	.09	-	.09	.13	**.14**	.10	.12	.07	.07	**.14**	.06	.09	.05
fi	.09	.07	.03	.09	-	.06	.08	.09	**.11**	.10	.09	.10	.07	**.13**	.06
fr	.07	.13	.10	**.13**	.06	-	**.15**	.08	.09	.06	.06	.09	.04	.07	.03
it	.08	.12	.08	**.14**	.08	**.15**	-	.09	.10	.07	.07	.11	.05	.08	.04
ja	.07	.08	.05	**.10**	.09	.08	.09	-	**.10**	.08	.07	.09	.05	.09	.08
nl	.10	.09	.05	**.12**	.11	.09	.10	.10	-	.10	.09	**.13**	.07	.12	.05
no	.08	.06	.03	.07	**.10**	.06	.07	.08	.10	-	.08	.09	.05	**.13**	.06
pl	.06	.06	.03	.07	.09	.06	.07	.07	**.09**	.08	-	.09	.08	**.09**	.05
pt	.10	.08	.05	**.14**	.10	.09	.11	.09	**.13**	.09	.09	-	.07	.11	.06
ru	.06	.04	.02	.06	**.07**	.04	.05	.05	.07	.05	**.08**	.07	-	.06	.05
sv	.09	.08	.04	.09	**.13**	.07	.08	.09	.12	**.13**	.09	.11	.06	-	.06
zh	.06	.03	.02	.05	**.06**	.03	.04	**.08**	.05	.06	.05	.06	.05	.06	-

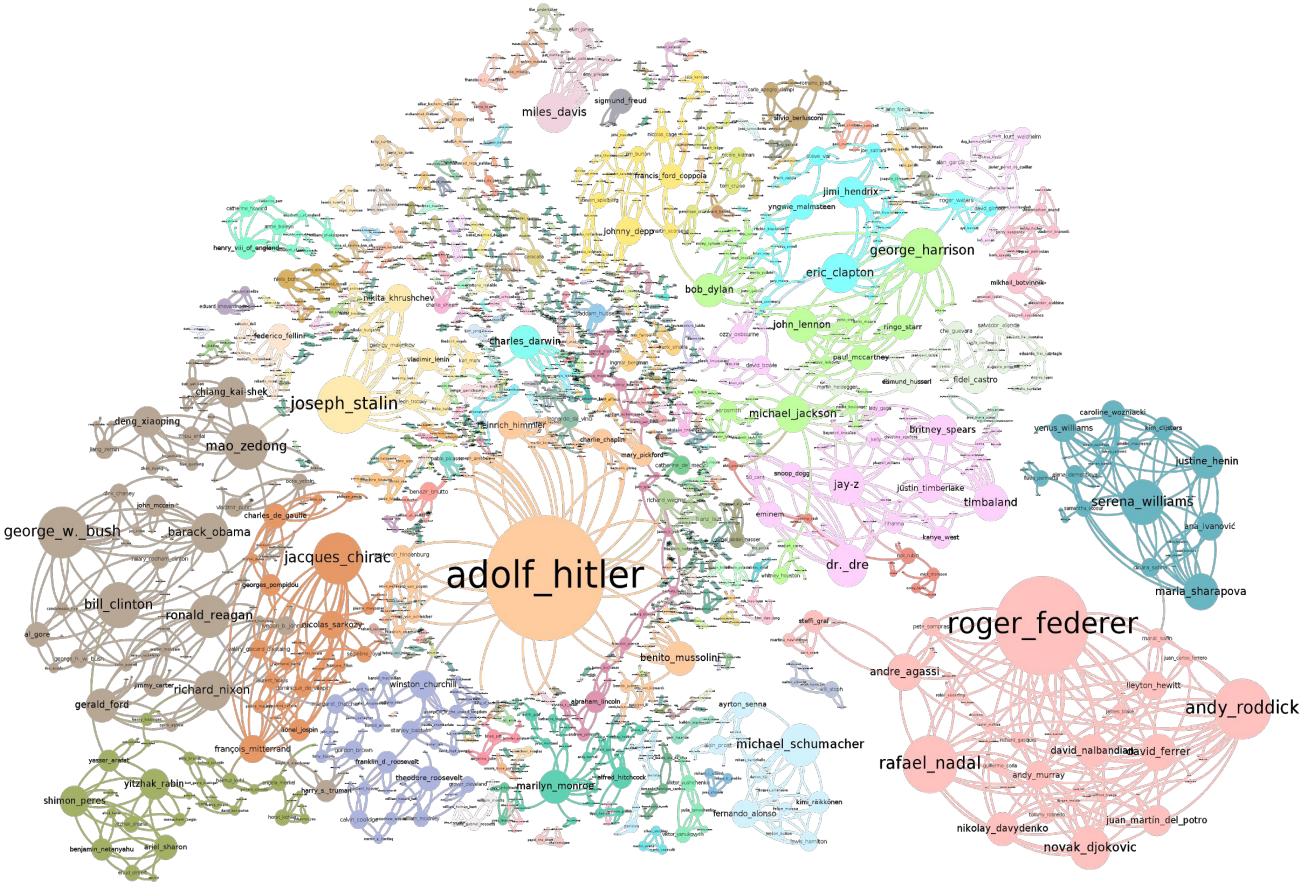
For each edition (row) the other two most similar versions are identified (also the most similar one is underlined)

Similarity among language editions



For each edition (node) we create links to the two most similar editions and then apply community detection

Intersection of networks in different languages



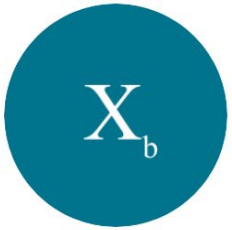
Conclusions

- Global social network measures are largely similar for all networks.
- Most central persons unveil interesting peculiarities about the language communities.
- Networks are more similar for geographically or linguistically closer communities.
- Many connections which can be found in most of the analysed language Wikipedias.

The World's Most Important People, According to Wikipedia

REBECCA J. ROSEN | APR 19 2012, 2:56 PM ET

The online manifestation of our collective cultural memory can give us a few clues to who we see as central figures.



IMPRECISE NEWS

A View from **Emerging Technology from the arXiv**

The Worrying Consequences of the Wikipedia Gender Gap

Male editors dramatically outnumber female ones on Wikipedia and that could be dramatically influencing the online encyclopedia's content, according to a new study

April 19, 2012

Wikipedia Editors Are Basically All Dudes - Should That Matter?

/ 04.30.2012 / Social Media

Wikipedia have released the results of their second editor survey and, much like the original one from April 2011, the vast majority of Wikipedia editors are still men.

Nobody will probably - and, really, should not - ever make the mistake of accusing Wikipedia of being perfect, but Wikipedia isn't even close to decent with respect to diversity. A new paper recently published by a team of researchers at the Barcelona Media Foundation in Spain found that the heavy majority of contributors being male on Wikipedia is producing a very concerning slant in the site's content. Examining the 15 largest language sites on Wikipedia, the researchers determined the top 5 most central persons from each language. Out of the 75 possible people listed, only 3 were women.

The Unz Review: An Alternative Media Selection

A Collection of Interesting, Important, and Controversial Perspectives Largely Excluded from the American Mainstream Media

← The Ted Nugent Threat & Other Stuff... Stereowiping and the Zimmerman Teletrav... →

iSteve Blog | April 2012 = 91 Posts | Current Post | Teasers

The Female Hitler Shortage & Other Great Moments in Feminist Theory

STEVE SAILER • APRIL 19, 2012 • 300 WORDS • 95 COMMENTS

← The Ted Nugent Threat & Other Stuff... Stereowiping and the Zimmerman Teletrav... →

Hide 95 Comments | Leave a Comment

95 Comments to "The Female Hitler Shortage & Other Great Moments in Feminist Theory"

Commenters to Ignore | ...to Follow | Endorsed Only | **Trim Comments?** No

1. **Ed** says:
   April 20, 2012 at 7:41 am GMT • 100 Words

   Buried in here is some fascinating data. I understand that the first four languages listed are English, French, German, and Italian, but I want to know what some of these other languages are. There is evidently a group of people, maybe Russians on the internet, who are fascinated by the lives of William Shakespeare, Napoleon II (!), Kenneth Branagh, and Elton John. Who are these people and how do their minds work?

   The other interesting question is that, given the ridiculously high unemployment in Spain, how "Pablo Aragon" (the name seems suspiciously made up) can get a paying job doing this, and where the "Barcelona Media Center" gets funding for this sort of work.

Case-study: Biographies on Wikipedia (II)

Eom, Y. H., Aragón, P., Laniado, D., Kaltenbrunner, A., Vigna, S., & Shepelyansky, D. L. (2015). Interactions of cultures and top people of Wikipedia from ranking of 24 language editions. PloS one, 10(3), e0114825.

Wikipedia

- The largest global knowledge repository
- The biggest collaborative platform on the Web
- 287 language editions (different descriptions of human knowledge)

Reflected cultural differences across language editions

Since language is one of the primary elements of culture, collective cultural biases may be reflected on the contents and organization of each Wikipedia language edition.

How can we identify reflected, collective, cultural difference across Wikipedia editions?

Methodology

1. For each edition → **Network of articles**

2. Ranking based on network structure
3. Identification of biographical articles and extraction of features
4. Analysis of top people

Edition	Articles	Edition	Articles
EN	4 212 493	**RU**	966 284
NL	1 144 615	**HE**	144 959
DE	1 532 978	**TR**	206 311
FR	1 352 825	**AR**	203 328
ES	974 025	**FA**	295 696
IT	1 017 953	**HI**	96 869
PT	758 227	**MS**	180 886
EL	82 563	**TH**	78 953
DA	175 228	**VI**	594 089
SV	780 872	**ZH**	663 485
PL	949 153	**KO**	231 959
HU	235 212	**JA**	852 087

The selection covers:

- ~ 60% of world population
- ~ 70% of Wikipedia articles

These 24 editions also cover languages which played an important role in human history including Western, Asian and Arabic cultures.

Methodology

1. For each edition → Network of articles

2. Ranking based on network structure
3. Identification of biographical articles and extraction of features
4. Analysis of top people

Seven Bridges of Königsberg

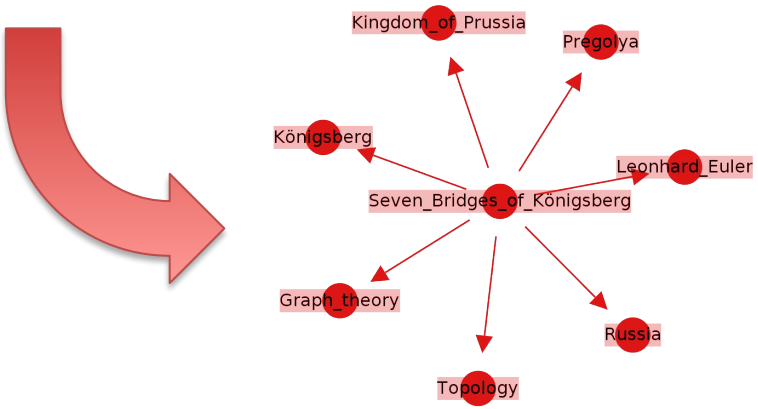
From Wikipedia, the free encyclopedia

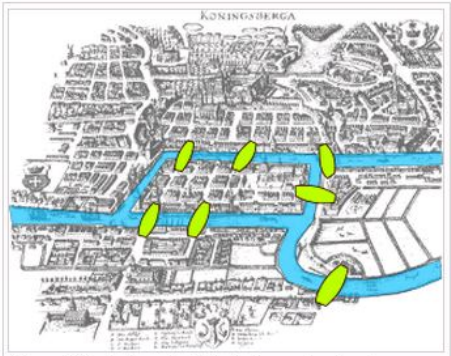
Coordinates: 54°42'12"N 20°30'56"E

The **Seven Bridges of Königsberg** is a historically notable problem in mathematics. Its negative resolution by Leonhard Euler in 1735 laid the foundations of graph theory and prefigured the idea of topology.[1]

The city of Königsberg in Prussia (now Kaliningrad, Russia) was set on both sides of the Pregel River, and included two large islands which were connected to each other and the mainland by seven bridges. The problem was to devise a walk through the city that would cross each bridge once and only once, with the provisos that: the islands could only be reached by the bridges and every bridge once accessed must be crossed to its other end. The starting and ending points of the walk need not be the same. The difficulty was the development of a technique of analysis and of subsequent tests that established this assertion with mathematical rigor.

Contents [hide]



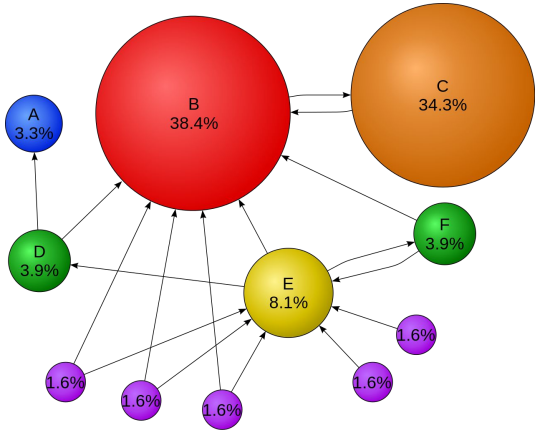


Map of Königsberg in Euler's time showing the actual layout of the seven bridges, highlighting the river Pregel and the bridges

Methodology

1. For each edition → Network of articles
2. **Ranking based on network structure**
3. Identification of biographical articles and extraction of features
4. Analysis of top people

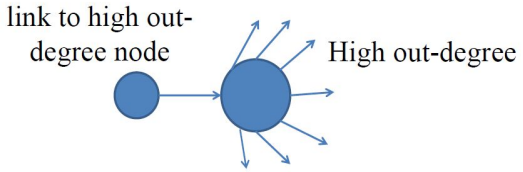
PAGERANK



$$P(i,t) = (1-\alpha)/N + \alpha \sum_j A_{ij} P(j,t-1)/k_{out}(j)$$

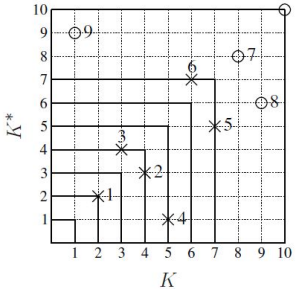
- Highly linked articles are important
- Articles cited by important articles are also important

CHEIRANK



$$P^*(i) = (1-\alpha)/N + \alpha \sum_j A_{ji} P^*(j)/k_{in}(j)$$

2DRANK

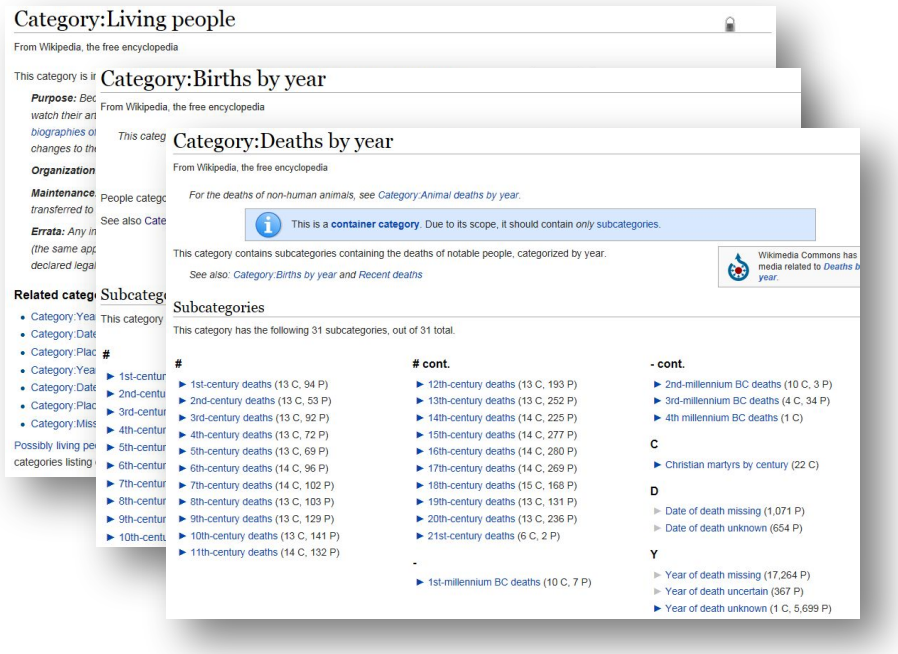


$$K'_i = max\{K_i, K^*_i\}$$

- Out-going links can be important

Methodology

1. For each edition → Network of articles
2. Ranking based on network structure
3. **Identification of biographical articles and extraction of features**
4. Analysis of top people



Extraction of features through DBpedia (manual inspection if missing):

- **Birth place** country level
- **Birth date** year level
- **Gender** male / female
- **Culture** current most spoken language at that birth place

- 1.1M people in EN-Wikipedia are identified
- Inter-language links: From EN to other language

Methodology

1. For each edition → Network of articles
2. Ranking based on network structure
3. Identification of biographical articles and extraction of features
4. **Analysis of top people**

EN-WIKIPEDIA NETWORK

Rank	PageRank persons	2DRank persons
1st	Napoleon	Frank Sinatra
2nd	Barack Obama	Michael Jackson
3rd	Carl Linnaeus	Pope Pius XII
4th	Elizabeth II	Elton John
5th	George W. Bush	Elizabeth II
6th	Jesus	Pope John Paul II
7th	Aristotle	Beyoncé Knowles
8th	William Shakespeare	Jorge Luis Borges
9th	Adolf Hitler	Mariah Carey
10th	Franklin D. Roosevelt	Vladimir Putin

GLOBAL NETWORK

Rank	PageRank global figures	Θ_{PR}	N_A	2DRank global figures	Θ_{2D}	N_A
1st	Carl Linnaeus	2284	24	Adolf Hitler	1557	20
2nd	Jesus	2282	24	Michael Jackson	1315	17
3rd	Aristotle	2237	24	Madonna (entertainer)	991	14
4th	Napoleon	2208	24	Jesus	943	14
5th	Adolf Hitler	2112	24	Ludwig van Beethoven	872	14
6th	Julius Caesar	1952	23	Wolfgang Amadeus Moza	853	11
7th	Plato	1949	24	Pope Benedict XVI	840	12
8th	William Shakespeare	1861	24	Alexander the Great	789	11
9th	Albert Eistein	1847	24	Charles Darwin	773	12
10th	Elizabeth II	1789	24	Barack Obama	754	16

Θ_A is the ranking score of algorithm A;
N_A is the number of appearances of a given person in the top100 rank for all editions.

Rank	PageRank global figures	Θ_{PR}	N_A
1st	Carl Linnaeus	2284	24
2nd	Jesus	2282	24
3rd	Aristotle	2237	24
4th	Napoleon	2208	24
5th	Adolf Hitler	2112	24
6th	Julius Caesar	1952	23
7th	Plato	1949	24
8th	William Shakespeare	1861	24
9th	Albert Eistein	1847	24
10th	Elizabeth II	1789	24

doi:10.1371/journal.pone.0114825.t004

"The reason for a somewhat unexpected PageRank leader Carl Linnaeus is related to the fact that **he laid the foundations for the modern biological naming scheme so that plenty of articles about animals, insects and plants point to the Wikipedia article about him**, which strongly increases the PageRank probability."

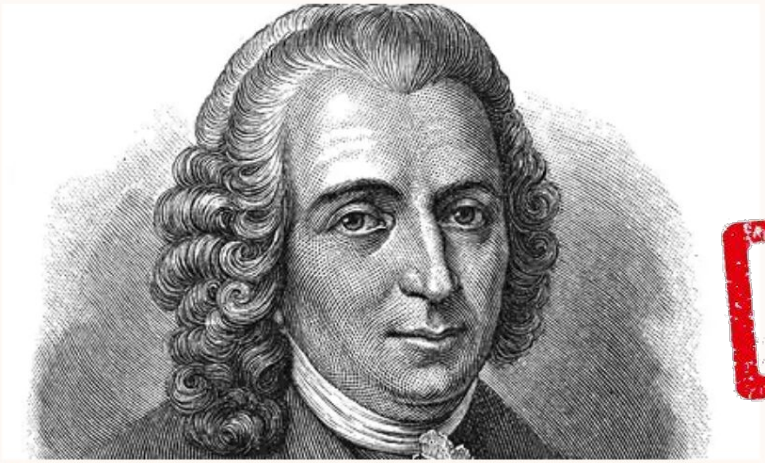
Taxonomy
Opinion

And the winner of Wikipedia's influence list is ... an 18th century botanist. Hear hear

Patrick Barkham

@patrick_barkham
Fri 13 Jun 2014 09.00 BST

Carl Linnaeus is hardly a household name, but the Swedish doctor who created a global naming system for species deserves this accolade



▲ 'Carl Linnaeus's great invention was the system of binomial nomenclature.' Photograph: Alamy

most popular


Live UK weather: 'beast from the east' hits Britain with freezing temperatures and snow – live


Arctic warming: scientists alarmed by 'crazy' temperature rises


Downing Street plays down Boris Johnson comments on Ireland


Live Brexit: Barnier says still 'significant points of disagreement' with UK on transition - Politics live


Harpist who performed for the Queen jailed for sexually abusing boy

FUN NEWS

SMARTNEWS *Keeping you current*

Who Does Wikipedia Think Is Bigger Than Jesus?

Swedish naturalist Carl Linnaeus is Wikipedia's most influential person, according to one measurement

Carl Linnaeus, 1707 to 1778. Swedish botanist, physician and zoologist. From Crabb's Historical Dictionary published 1825. (Ken Welsh/Design Pics/Corbis)

15

DAILY NEWS 10 June 2014

Jesus and Hitler beaten in Wikipedia influence list

By Jacob Aron

The who's who of Wikipedia: (top) Adolf Hitler, Carl Linneaus, Jesus, Barack Obama; (bottom) Pope Pius XII, Frank Sinatra, Michael Jackson, Napoleon *(Images: Rex)*

Carl Linnaeus ranked most influential person of all time

13 June 2014

NATIONALIST NEWS

Researchers at the University of Toulouse have used Wikipedia and network theory to rank the world's most influential people through history. At the top of the list we find none other than Uppsala's own Carl Linnaeus.


THE SCIENCE NEWS CYCLE
JORGE CHAM © 2009
Start Here
Your Research
Conclusion: A is correlated with B (ρ=0.56), given C, assuming D and under E conditions.
...is translated by...
UNIVERSITY PR OFFICE
(YES, YOU HAVE ONE)
FOR IMMEDIATE RELEASE: SCIENTISTS FIND POTENTIAL LINK BETWEEN A AND B (UNDER CERTAIN CONDITIONS).
...which is then picked up by...
NEWS WIRE ORGANIZATIONS
A CAUSES B, SAY SCIENTISTS.
...who are read by ...
THE INTERNETS
Scientists out to kill us again.
POSTED BY RANDOM DUDE
Comments (377)
OMG! i kneeew it!!!
WTH???????
...then noticed by...
We saw it on a Blog!
A causes B all the time
What will this mean for Obama?
BREAKING NEWS BREAKING NEWS BREA
...and caught on ...
CNC Cable NEWS
A=B
LOCAL EYEWITLESS NEWS
...eventually making it to...
WHAT YOU DON'T KNOW ABOUT "A"... CAN KILL YOU! MORE AT 11...
A: KILLER AMONG US??
LOCAL EYEWITLESS NEWS
I'M WEARING THIS TO WARD OFF "A"
YOUR GRANDMA
WWW.PHDCOMICS.COM

Top people ranking

NewScientist Tech

Jesus and Hitler beaten in Wikipedia influence list

And the winner of Wikipedia's influence list is ... an 18th century botanist. Hear hear Patrick Barkham

theguardian
Winner of the Pulitzer prize 2014

SMARTNEWS *Keeping you current*

Who Does Wikipedia Think Is Bigger Than Jesus?

Swedish naturalist Carl Linnaeus is Wikipedia's most influential person, according to one measurement

THE INDEPENDENT TUESDAY 26 MAY 2015

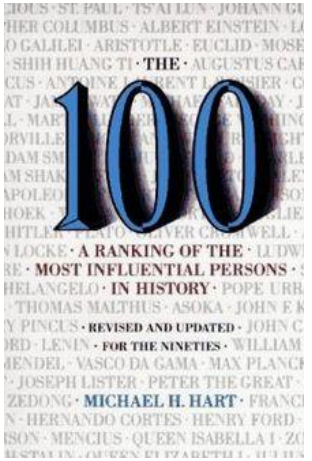
Which 18th century botanist is more influential than both Jesus and Hitler on Wikipedia?

Not bigger than Jesus, but relevant from an encyclopedic point of view

Overlap with alternative rankings:

- Hart's Influential personas in History: **43%**
- Top 100 people Pantheon MIT list: **44%**

PANTHEON
MAPPING HISTORICAL CULTURAL PRODUCTION



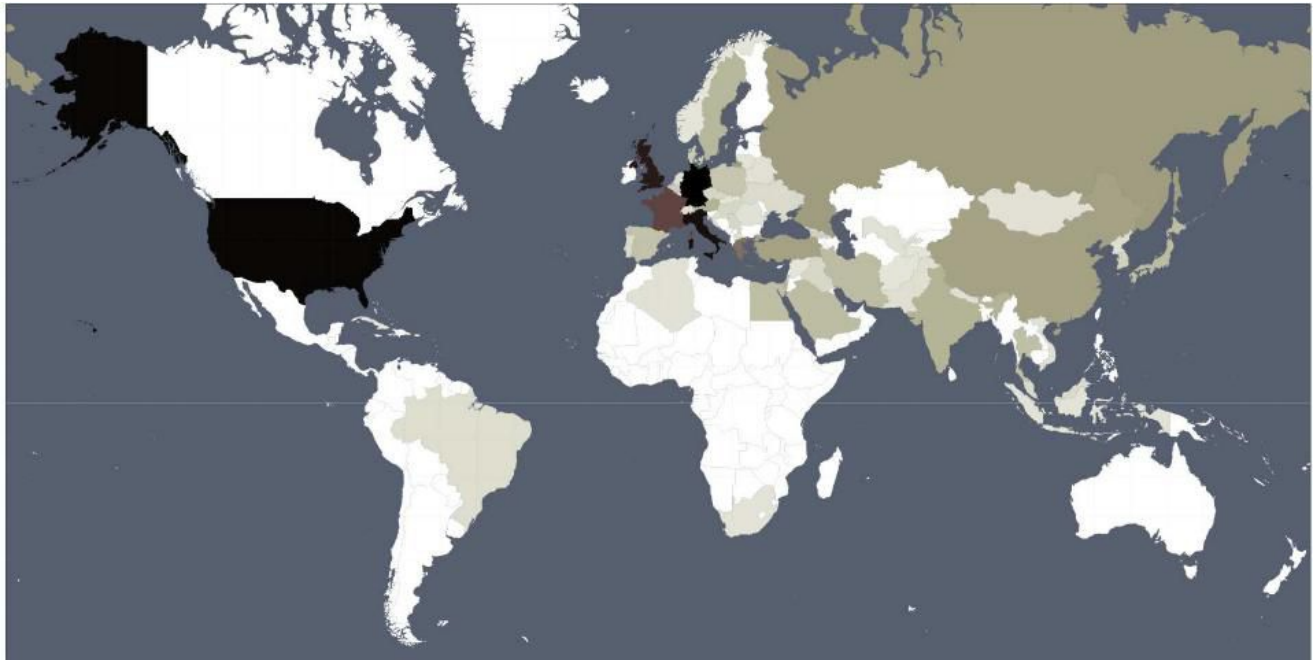
Assessing cultural differences in top 100 people

SPATIAL DISTRIBUTION

TEMPORAL DISTRIBUTION

GENDER DISTRIBUTION

Results: Spatial distribution (I)

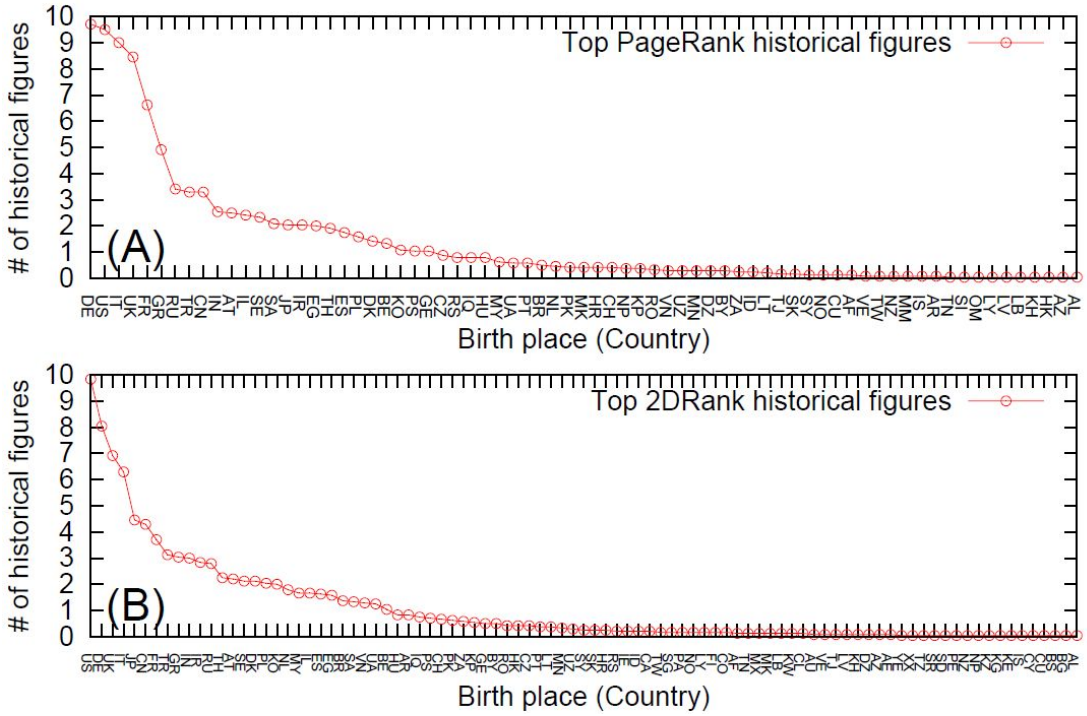


Sum of appearances of historical figures from a given country in the 24 lists of top 100 persons for PageRank.
Color changes from zero (white) to 233 for Germany (black).

Overall Western

11 editions are not European (AR, FA, HE, HI, JA, KO, MS, TH, TR,VI, ZH)

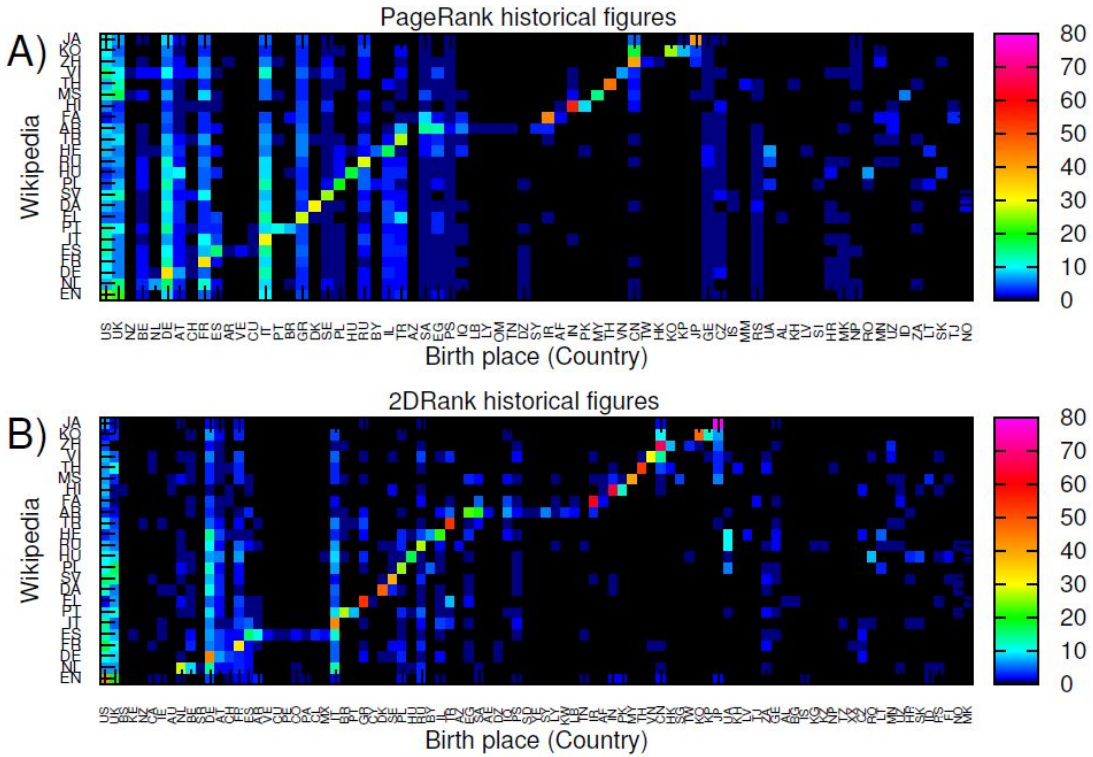
Results: Spatial distribution (II)



Birth place distribution of top historical figures averaged over 24 Wikipedia edition for (A) PageRank historical figures (71 countries) and (B) 2DRank historical figures (91 countries).

Overall Western

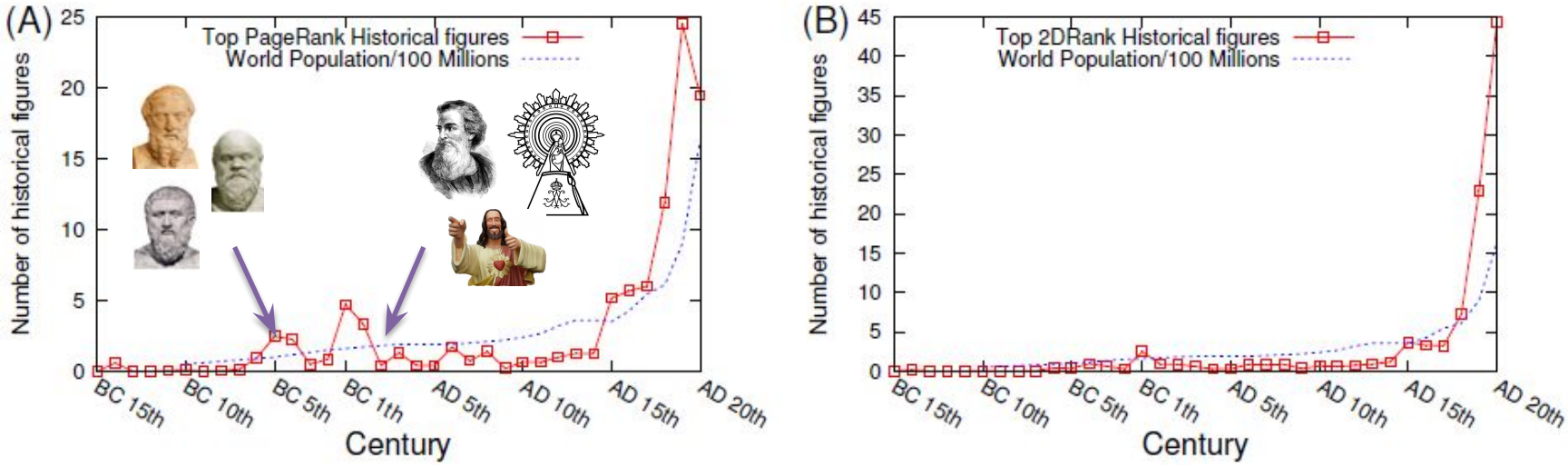
Results: Spatial distribution (III)



For each Wikipedia edition, distributions of (A) PageRank historical figures over 71 countries;
(B) 2DRank historical figures over 91 countries

Local skewness in the spatial distribution of the top historical figures

Results: Temporal distribution (I)

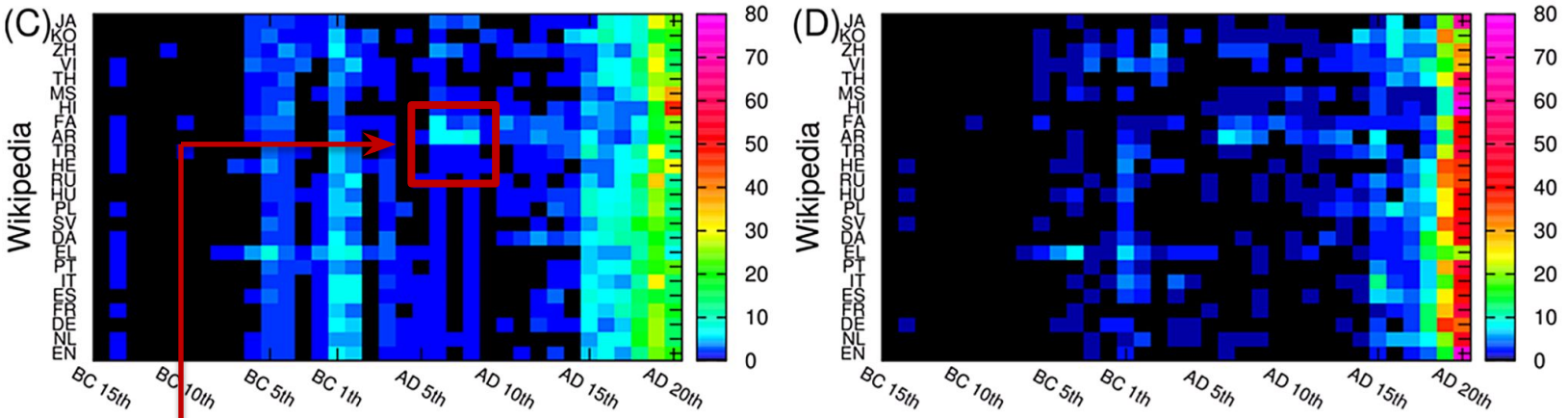


Birth date distribution of (A) PageRank historical figures (B) 2DRank historical figures.

Most are born after 17th century.

Peaks in BC 5th and BC 1st century in top PageRank people

Results: Temporal distribution (II)

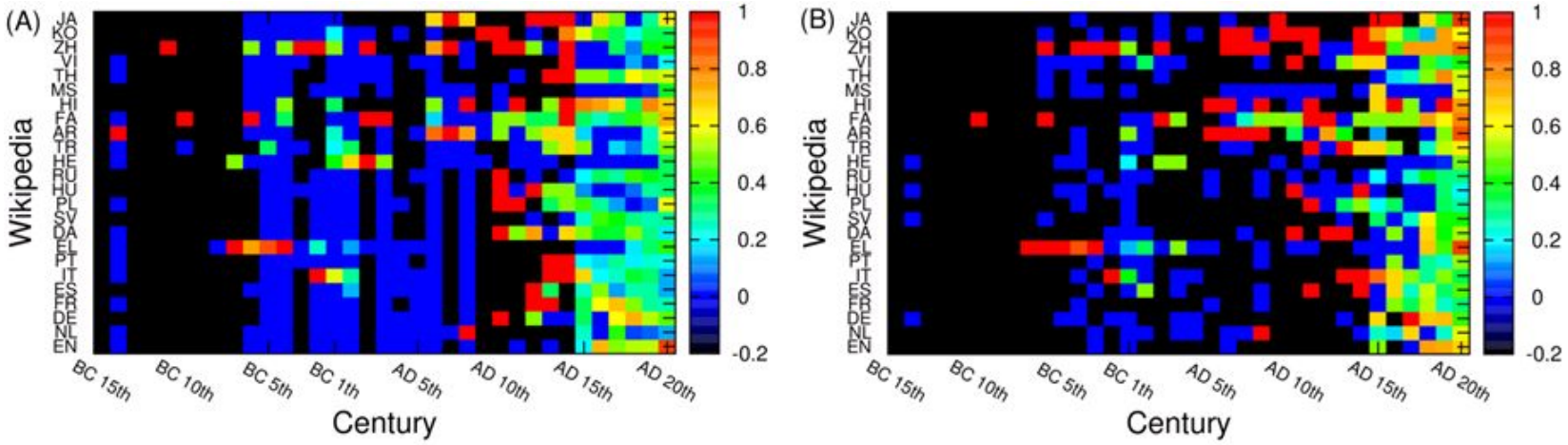


For each Wikipedia edition, birth date distributions of (C) PageRank historical figures (D) 2DRank historical figures.

Most are born after 17th century

Arabic & Persian from 6th to 10th century

Results: Spatial & Temporal distribution



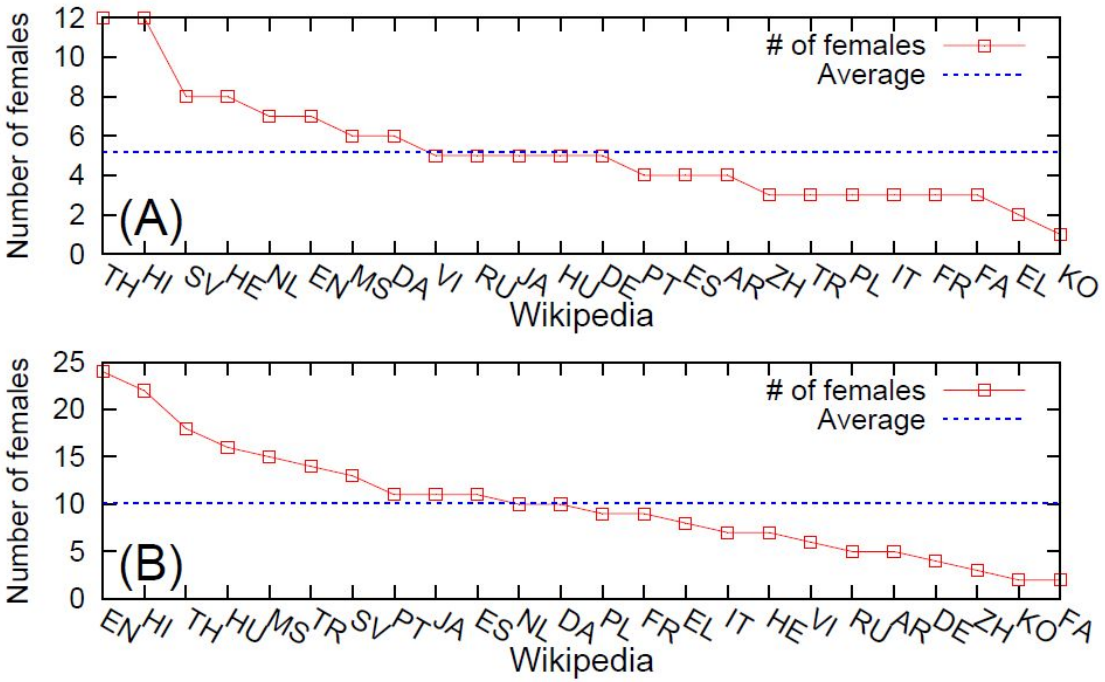
Temporal locality property of cultures: $r_{L,C} = M_{L,C} / N_{L,C}$; of (A) PageRank historical figures (B) 2DRank historical figures.

$M_{L,C}$ is the number of historical figures born in countries attributed to a given language edition L

$N_{L,C}$ is the total number of historical figures in a given edition L

- **Locality in Ancient History:** Greek, Italian, Hebrew or Chinese editions
- **Locality in Middle Ages:** French, Spanish or Portuguese editions
- **Locality in Modern History:** English edition

Results: Gender distribution (I)

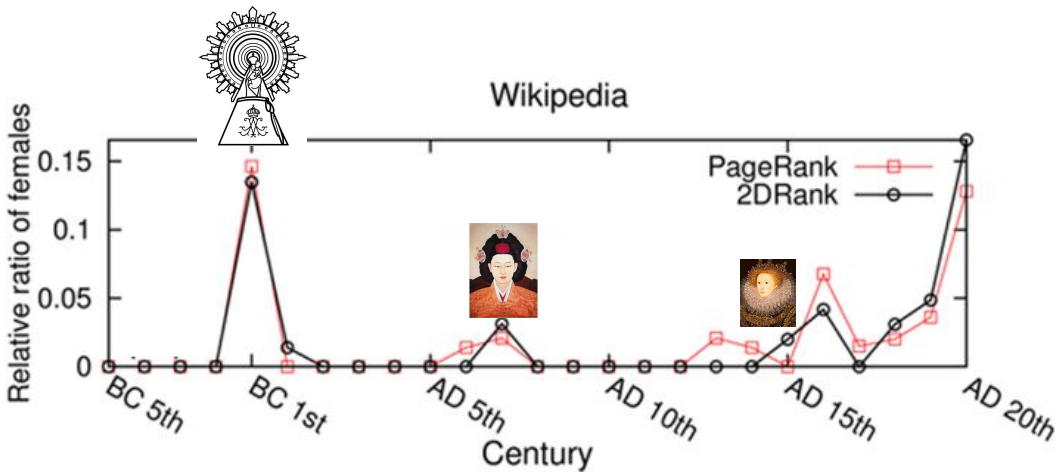


For each Wikipedia edition, number of females of
(A) Top PageRank historical figures;
(B) Top 2DRank historical figures.

Overall male skewed

Rank	TH PageRank
1	**Sirindhorn**
2	Bhumibol Adulyadej
3	**Sirikit**
4	Thaksin Shinawatra
5	Gautama Buddha
6	Adolf Hitler
7	Taksin
8	**Queen Victoria**
9	Abhisit Vejjajiva
10	Pridi Banomyong
11	**Yingluck Shinawatra**
12	**Galyani Vadhana**
13	**Srinagarindra**
14	J. R. R. Tolkien
15	Samak Sundaravej

Results: Gender distribution (II)



The average female ratio of historical figures in given centuries across 24 Wikipedia editions.

The ratio of women has grown in the last centuries

Rank	Θ_{PR}	N_A	PageRank female figures	*CC*	Century	*LC*
1	1789	24	Elizabeth II	UK	20	EN
2	1094	17	Mary (mother of Jesus)	IL	-1	HE
3	404	12	Queen Victoria	UK	19	EN
4	234	6	Elizabeth I of England	UK	16	EN
5	128	2	Maria Theresa	AT	18	DE
6	100	1	Benazir Bhutto	PK	20	HI
7	94	1	Catherine the Great	PL	18	PL
8	91	1	Anne Frank	DE	20	DE
9	87	1	Indira Gandhi	IN	20	HI
10	86	1	Margrethe II of Denmark	DK	20	DA

Rank	Θ_{2D}	N_A	2DRank female figures	*CC*	Century	*LC*
1	991	14	Madonna (entertainer)	US	20	EN
2	664	9	Elizabeth II	UK	20	EN
3	580	8	Mary (mother of Jesus)	IL	-1	HE
4	550	9	Queen Victoria	UK	19	EN
5	225	5	Agatha Christie	UK	19	EN
6	211	4	Mariah Carey	US	20	EN
7	206	7	Britney Spears	US	20	EN
8	200	3	Margaret Thatcher	UK	20	EN
9	191	2	Martina Navratilova	CZ	20	WR
10	175	2	Elizabeth I of England	UK	16	EN

Results: Networks of cultures

Cross cultural top persons as influence between cultures

Ex)

EN edition

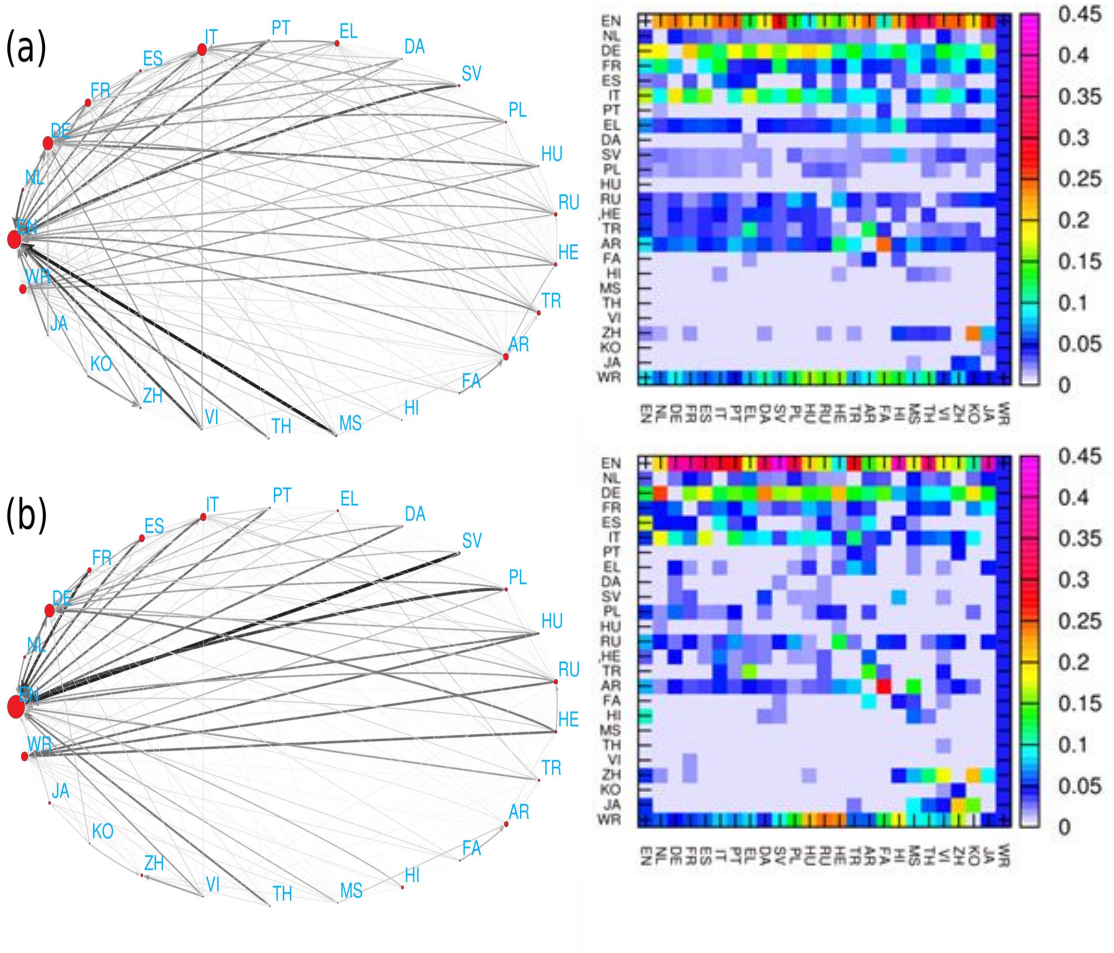
Napoleon
Napoleon III
Descartes

3

2

FR edition

Shakespeare
Elizabeth II



Network of cultures and corresponding Google Matrix obtained from 24 Wikipedia languages and the remaining world (WR) considering (A) top PageRank historical figures and (B) 2DRank historical figures.

Summary

Global: Most important historical figures across Wikipedia language editions were:

- from Western countries
- born after the 17th century
- male

Local: Each edition highlights historical figures from that culture

Future work

Investigate the origins of skewness (description → explanation)

Refine the extraction of features through WikiData project (instead of DBpedia)

Compare results to alternative categories of Wikipedia (e.g. food)

Case-study: Barcelona 2016

Aragón, P., Gallego, H., Laniado, D., Volkovich, Y., & Kaltenbrunner, A. (2017). Online network organization of Barcelona en Comú, an emergent movement-party. Computational Social Networks. doi:10.1186/s40649-017-0044-4

https://computationalsocialnetworks.springeropen.com/articles/10.1186/s40649-017-0044-4

Movement organization

Networked social movement: Networked in multiple forms (multimodal, on/offline, across platforms) without a central node, and with a decentered structure

(Castells, 2013)

Change from logic of collective action to a logic of **connective action**

(Bennett & Segerberg, 2013)

15M

#SPANISH REVOLUTION

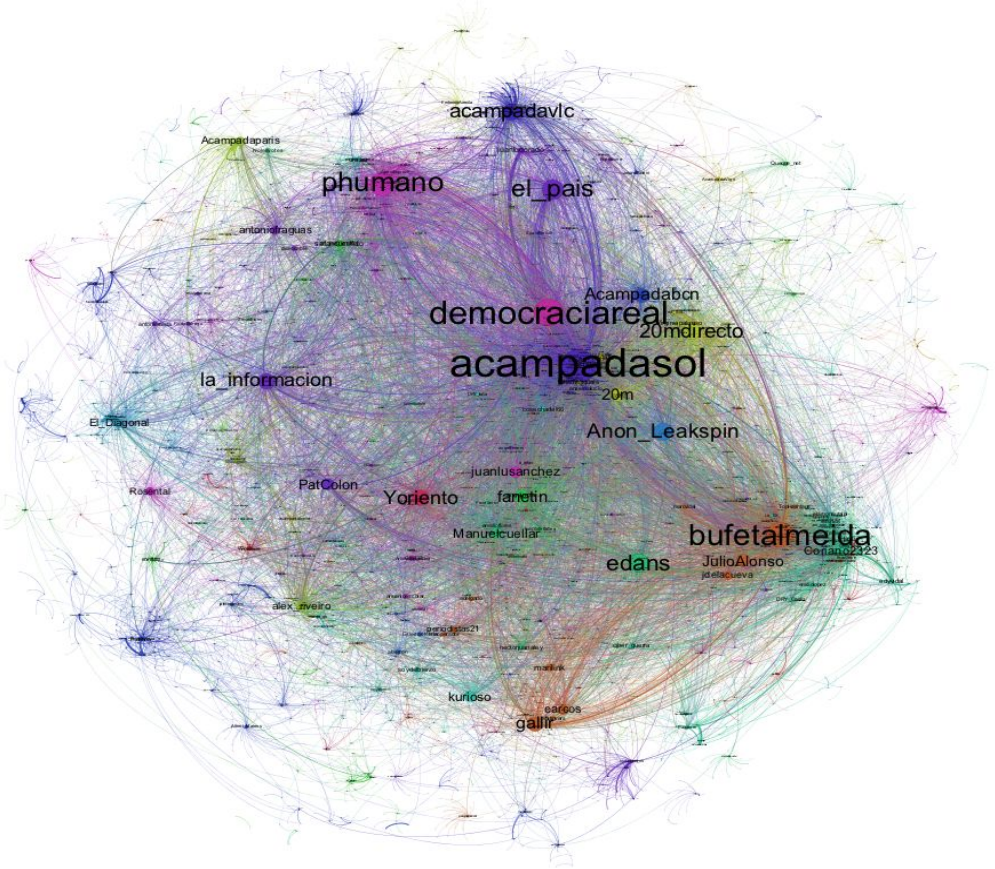
Movement networks

“Decentralized structure, based on coalitions of smaller organizations”

(González-Bailón et al, 2011)

“Decentralized organization, without leaders or stable representatives”

(Aragón et al, 2015)



RT network of the 15M movement
May 15-22, 2011 (Aragón et al, 2015)

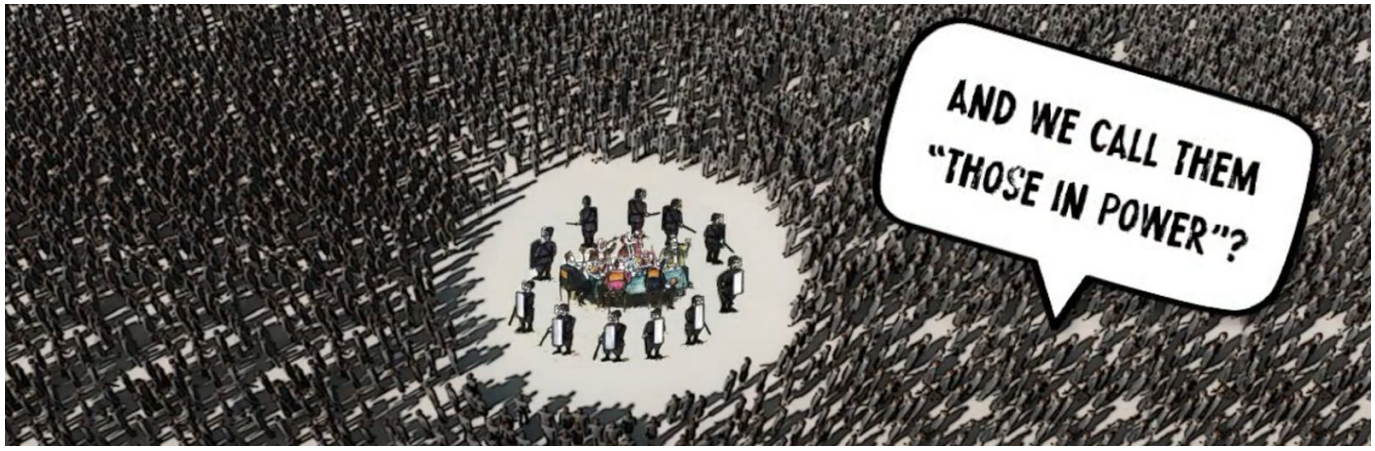
Party organization

Iron Law of Oligarchy: Political parties, like any complex organization, self-generate an elite ("Who says organization, says oligarchy")

(Michels, 1915)

Elite theory: Small minorities (elites) hold the most power in political processes

(Pareto et al, 1935; Mosca, 1939; Mills 1999)

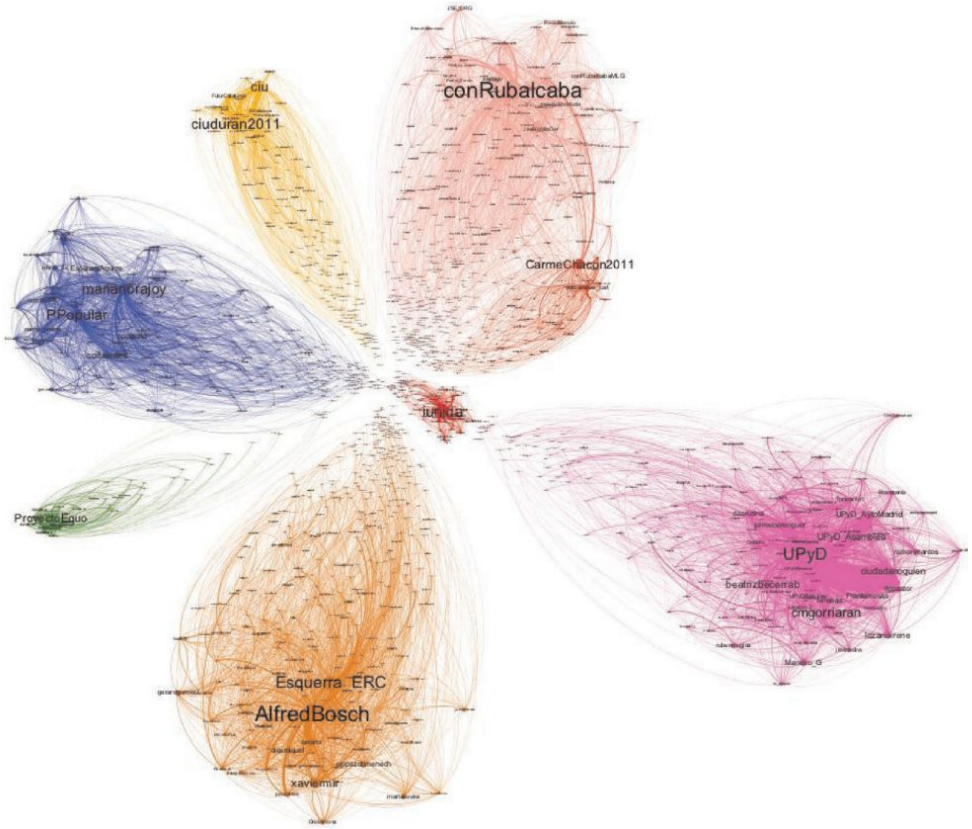


Party networks

The Twitter party networks in the 2011 Spanish election presented:

- Isolated clusters for each party.
- Minor and new parties were more clustered and better connected.
- Every party cluster was **strongly centralized** around candidate and/or party profiles.

(Aragón et al, 2013)



RT network of political parties in the 2011 Spanish election (Aragón et al, 2013)

Spanish Local Elections 2015

Grassroots parties emerged from the 15M movement:

- Barcelona en Comú
- Ahora Madrid
- Zaragoza en Común
- Marea Atlántica
- Compostela Aberta
- Por Cádiz Sí se puede
- Guanyem Badalona en Comú

Research question

Assuming that:

- **Barcelona en Comú** emerged from the 15M movement
- the **15M movement** followed a decentralized structure

Has Barcelona en Comú…

- preserved a decentralized structure?
- adopted a conventional centralized organization?

Research question

"Political parties share some interesting patterns of behavior, but also exhibit some unique and interesting idiosyncrasies" (e.g. tagging practice of politicians)

(Lietz et al, 2015)

Political Party / Coalition	Party account(s)	Candidate account
CiU - Convergència i Unió	@CDCBarcelona @uniobcn	@xaviertrias
PSC - Partit dels Socialistes de Catalunya	@pscbarcelona	@jaumecollboni
PP - Partit Popular de Catalunya	@PPBarcelona_	@albertofdezxbcn
BeC - Barcelona en Comú	@bcnencomu @icveuiaBCN @Podem_BCN @Equobcn @pconstituentBCN	@AdaColau
ERC - Esquerra Republicana de Catalunya	@ERCbcn	@AlfredBosch
Cs - Ciudadanos	@Cs_BCNA	@CarinaMejias
CUP - Capgirem Barcelona	@CapgiremBCN @CUPBarcelona	@MJLecha

Sampling criteria based on candidate and party accounts:

373 818 RTs

RT network

- 6 492 nodes
- 16 775 edges

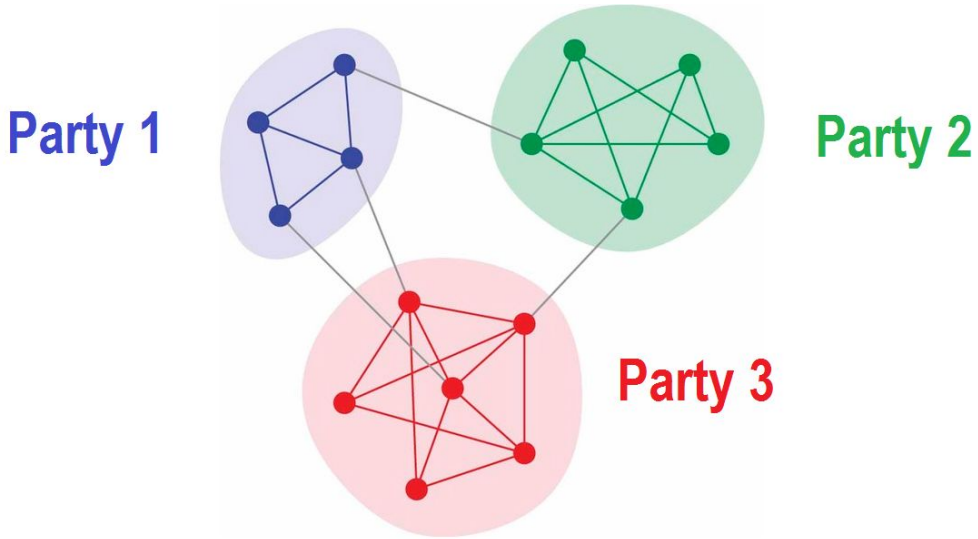
Methodology

Community detection

Identify the organization of nodes in clusters: political party networks.

Cluster characterization

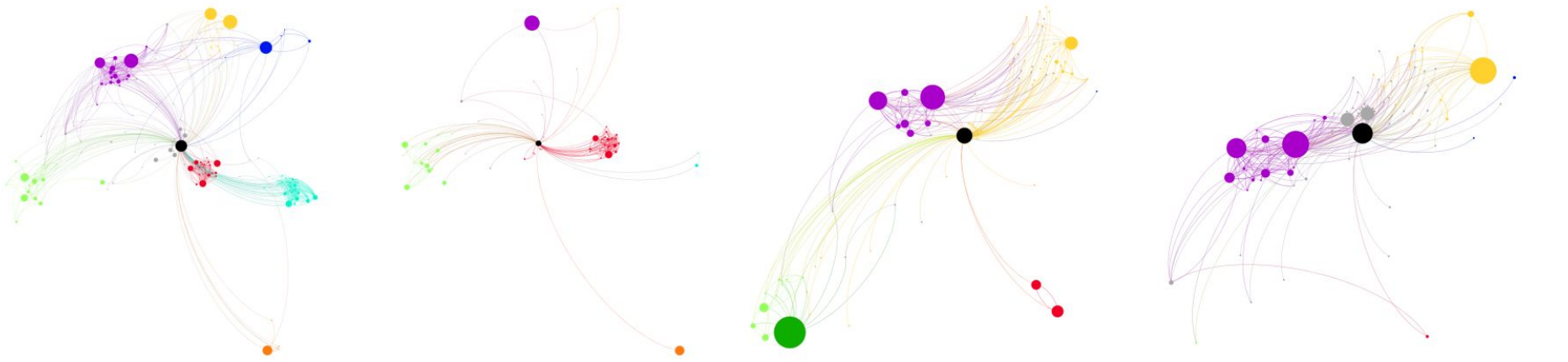
Characterize the topology of the intra-network of each cluster.



Community detection

First result with the Louvain Method (Blondel et al, 2008):

- Eight major clusters (seven parties)
- Every cluster contains some media accounts: media build weak ties
- Analysis of the ego-network of relevant media accounts:
  - Public TV account retweeted by users from every cluster
  - Private media mostly retweeted by users from like-minded parties



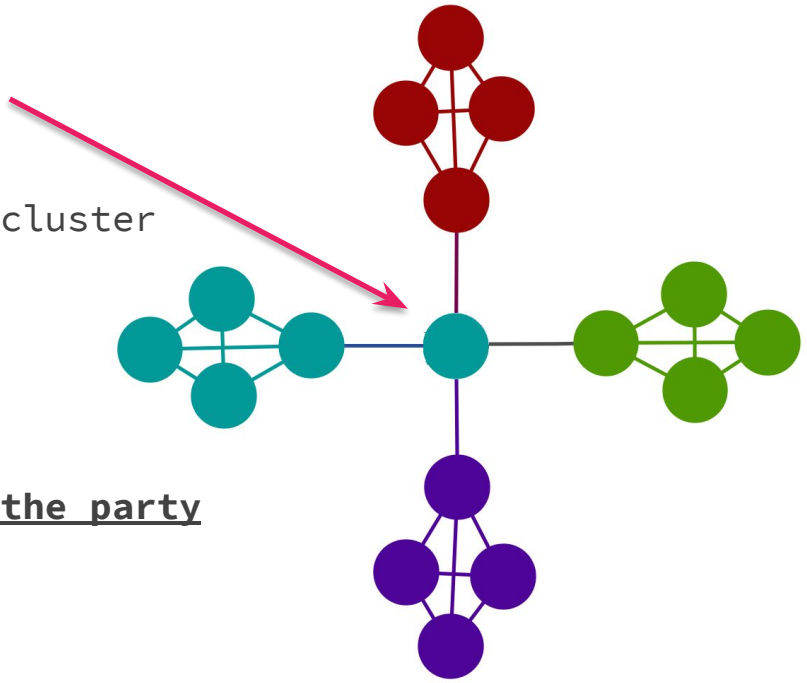
(a) @btvnoticies (public media) (b) @elpaiscat (private progressive media) (c) @arapolitica (private Catalan nationalist media) (d) @naciodigital (private Catalan nationalist media)

Community detection

Every cluster contains some media accounts

Each execution produces different results:

Some media do not always belong to the same cluster



We want the real intra-network structure of the party

Confident version of the Louvain Method

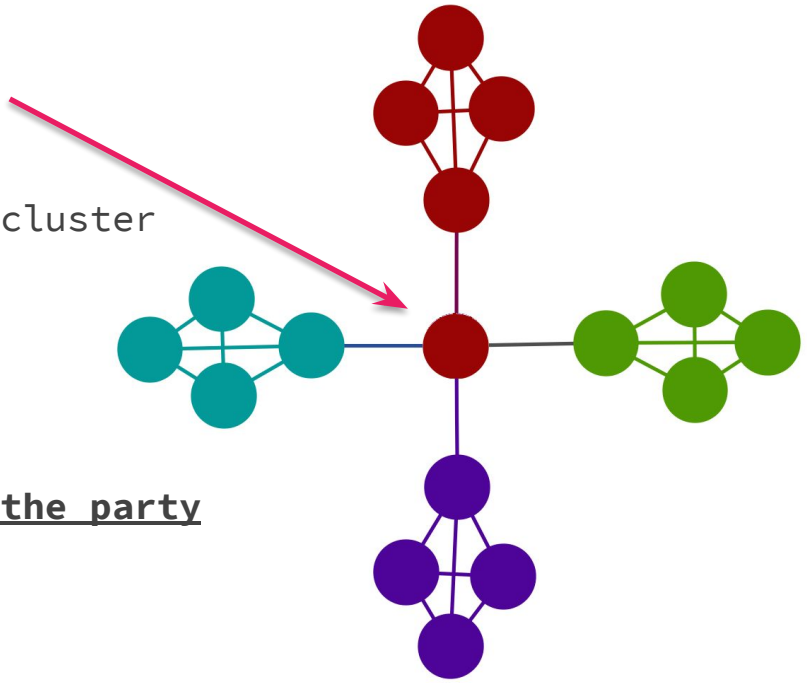
- Multiple executions (N=100)
- Identify each cluster by the most relevant user (PageRank)
- Just consider nodes that appear in the same cluster many times (1-ε=0.95)

Community detection

Every cluster contains some media accounts

Each execution produces different results:

Some media do not always belong to the same cluster



We want the real intra-network structure of the party

Confident version of the Louvain Method

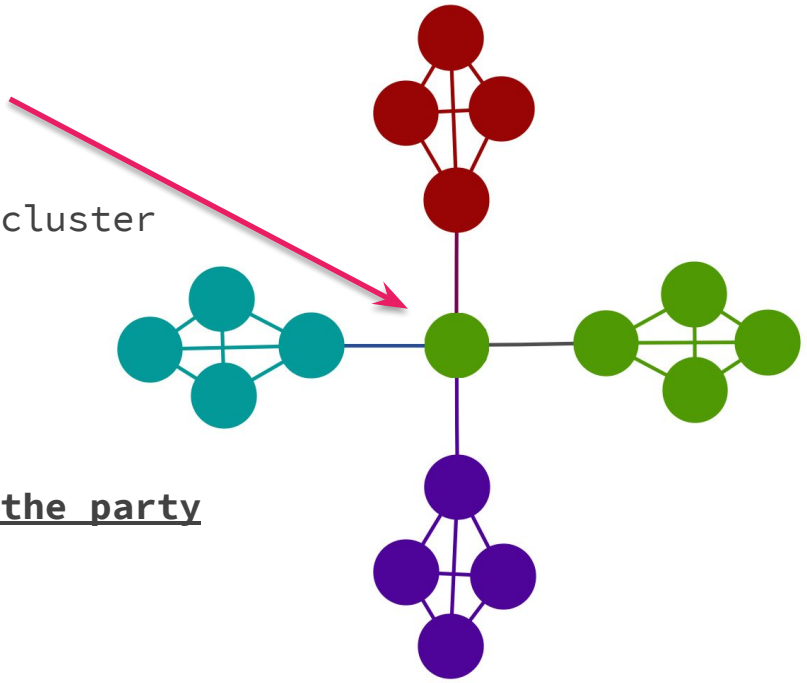
- Multiple executions (N=100)
- Identify each cluster by the most relevant user (PageRank)
- Just consider nodes that appear in the same cluster many times (1-ε=0.95)

Community detection

Every cluster contains some media accounts

Each execution produces different results:

Some media do not always belong to the same cluster



We want the real intra-network structure of the party

Confident version of the Louvain Method

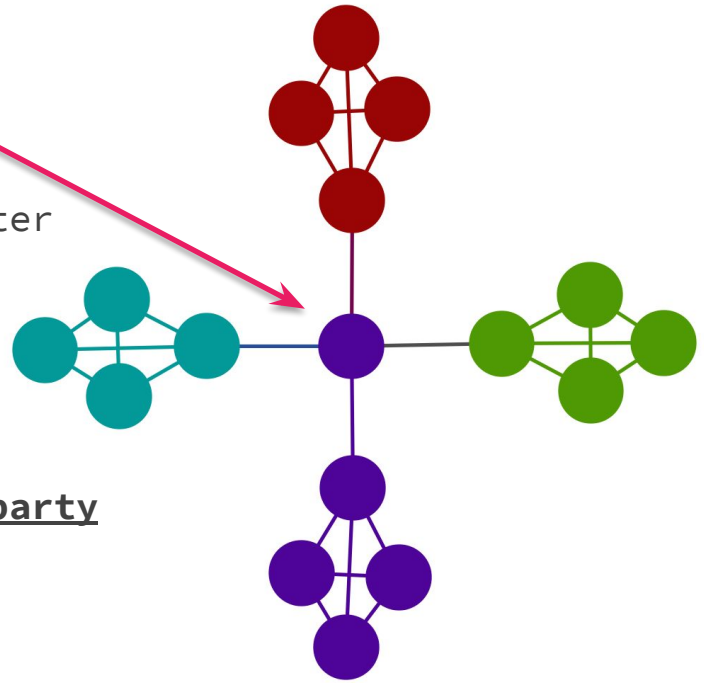
- Multiple executions (N=100)
- Identify each cluster by the most relevant user (PageRank)
- Just consider nodes that appear in the same cluster many times (1-ε=0.95)

Community detection

Every cluster contains some media accounts

Each execution produces different results:

Some media do not always belong to the same cluster



We want the real intra-network structure of the party

Confident version of the Louvain Method

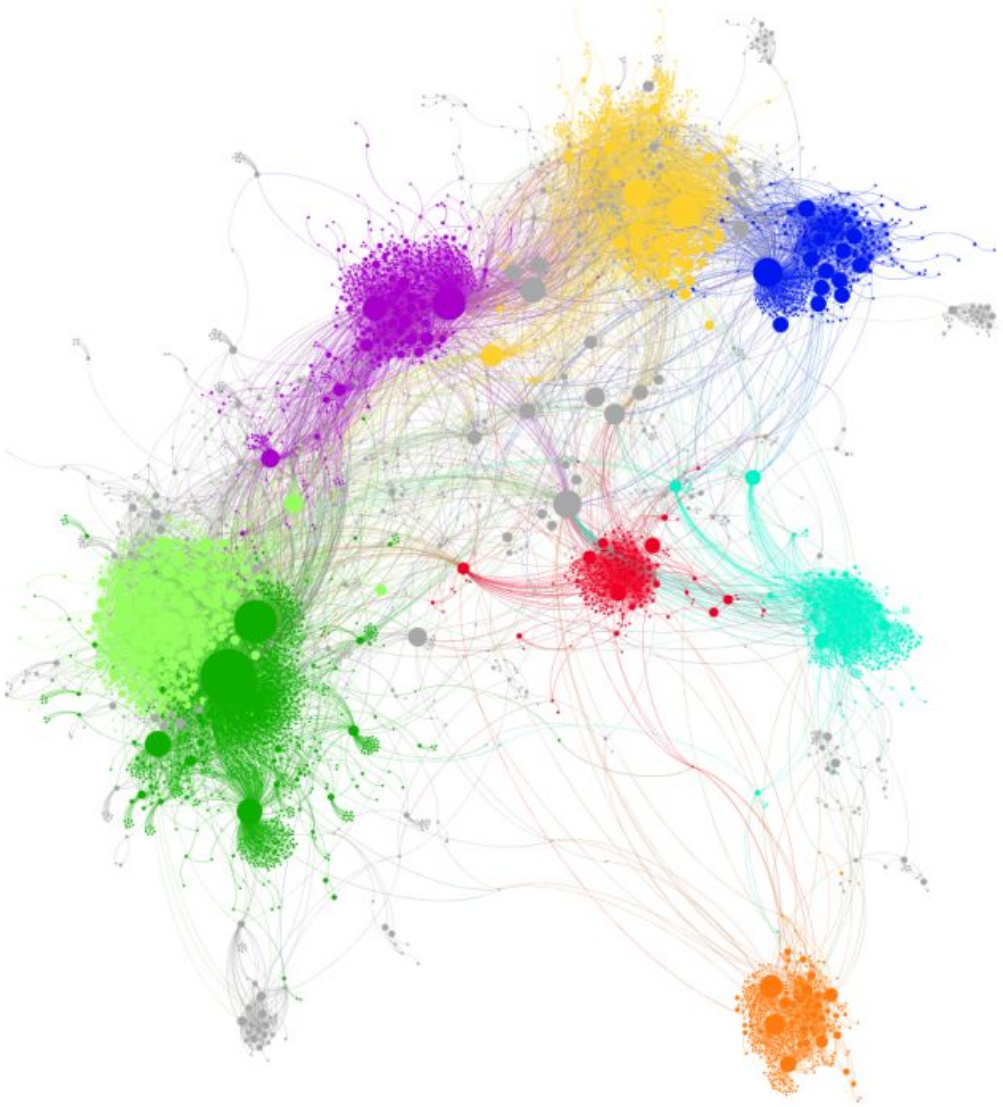
- Multiple executions (N=100)
- Identify each cluster by the most relevant user (PageRank)
- Just consider nodes that appear in the same cluster many times (1-ε=0.95)

Community detection

Results with the **extended version of the Louvain Method**

Constant presence of eight major clusters (seven parties) along the 100 executions:

- Most media accounts do not appear in major clusters
- Two clusters for Barcelona en Comú



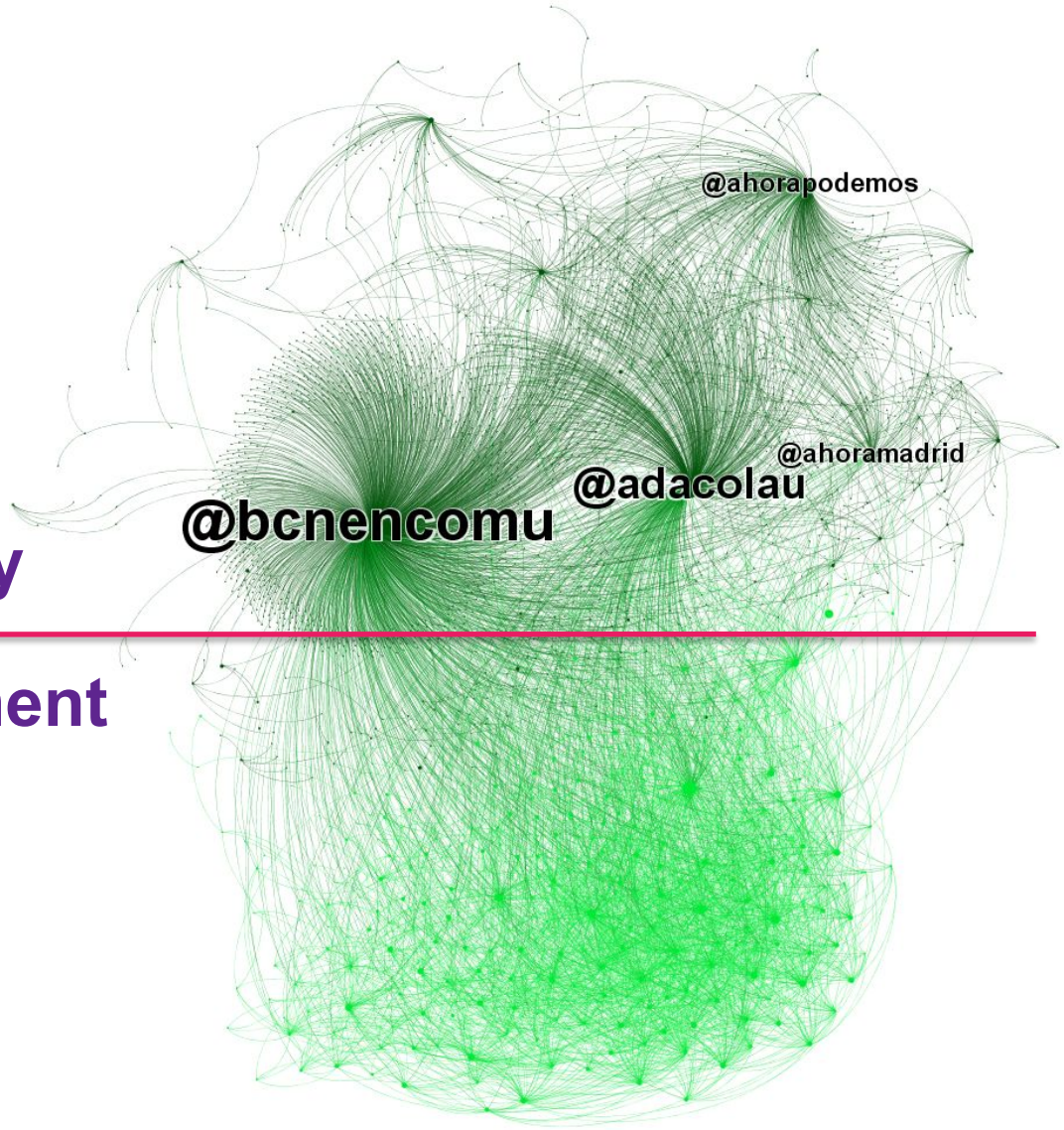
Community detection

Results with the **extended version of the Louvain Method**

Constant presence of eight major clusters (seven parties) along the 100 executions:

- Most media accounts do not appear in major clusters
- Two clusters for Barcelona en Comú

Community detection

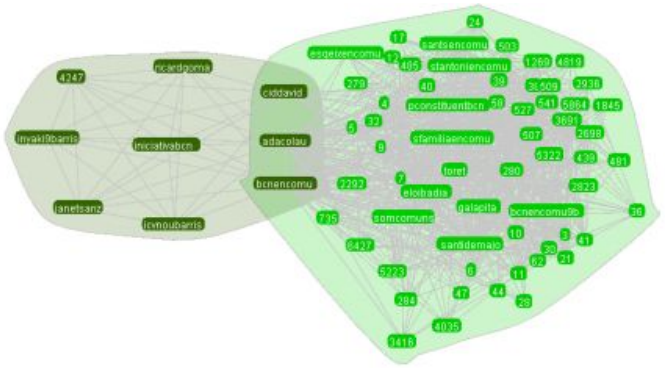


Community detection

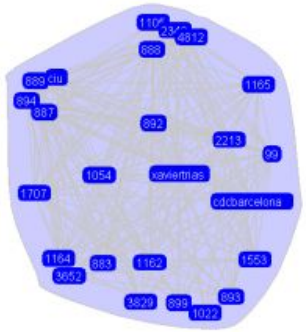
Clique Percolation Model

Similar results but…

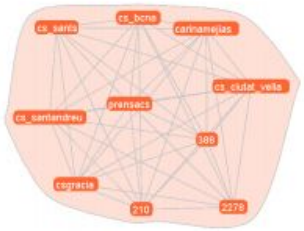
- CPM is O(exp(n))
  (NP-complete problem)
- CPM is not sensitive to different sizes and structures of parties
- K-cliques are only the core of the structure of party networks
  (the periphery is relevant)



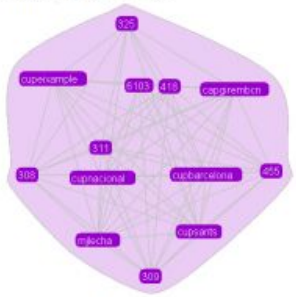
(a) BeC-p (left) & BeC-p (right) (k=9).



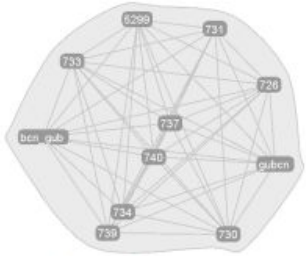
(b) CiU ($k = 9$).



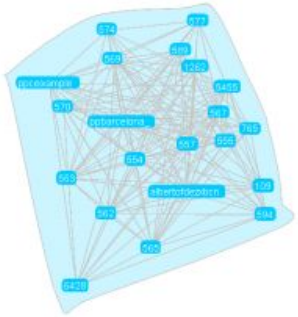
(c) Cs ($k = 9$).



(d) CUP ($k = 9$).

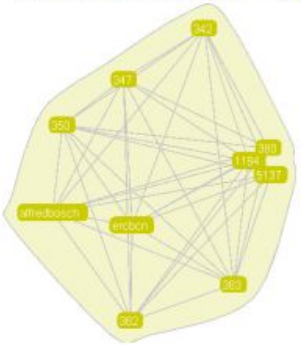


(e) Guardia Urbana ($k = 9$).



(f) PP (k=9).

(g) PSC ($k = 8$).



(h) ERC ($k = 7$).

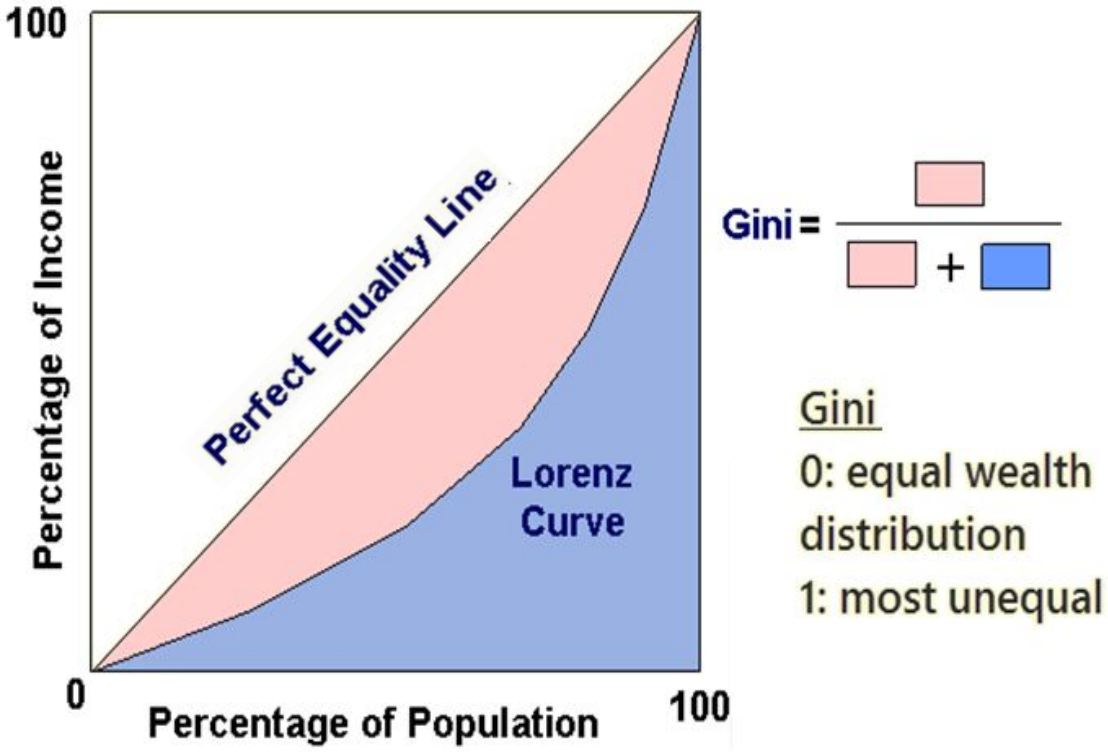
Cluster characterization

Inspired by the social dimensions of García et al. (2015):

- Hierarchical structure
  In-degree centralization → Gini coefficient of the in-degree distribution
- Small world phenomenon (f.k.a. information efficiency)
  Avg. path length + Clustering coefficient
- Coreness (f.k.a. social resilience)
  Maximal k-core → Distribution of k-indices

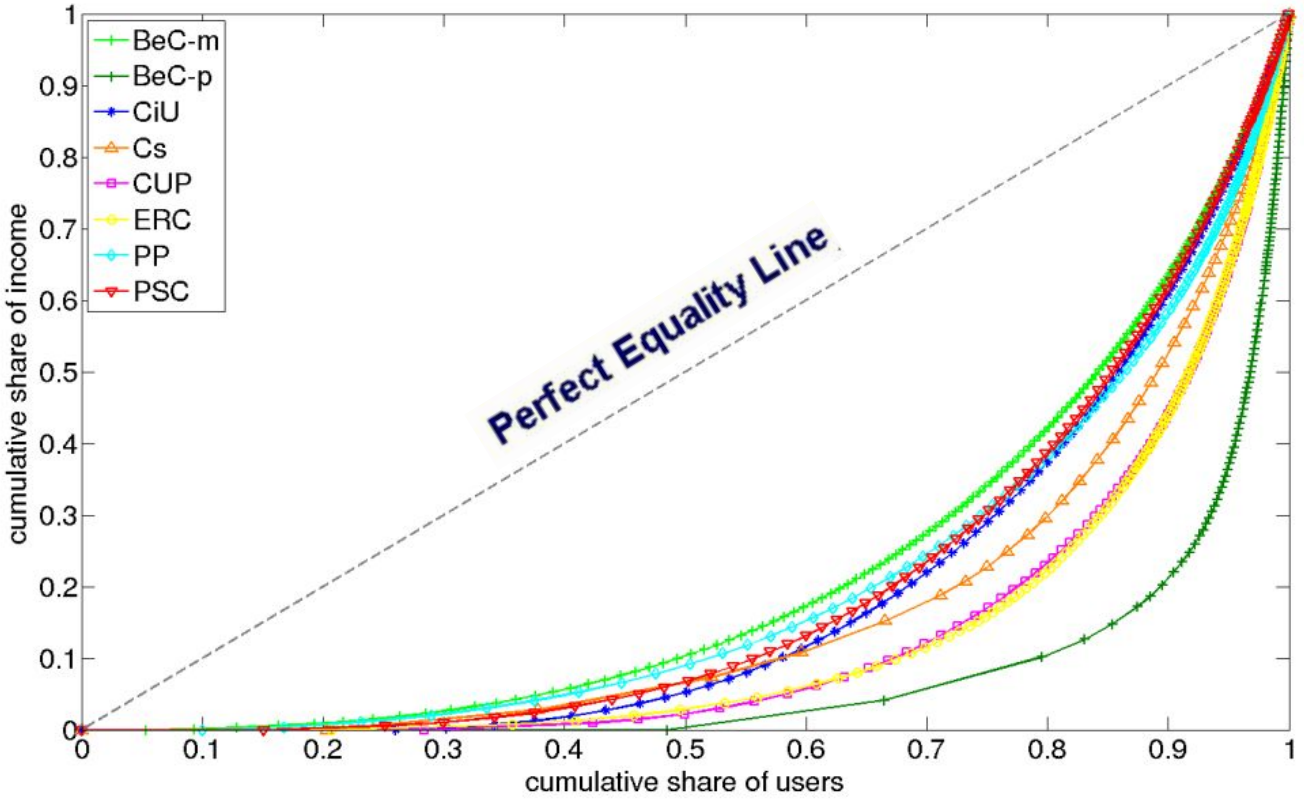
Hierarchical structure

Gini coefficient of the in-degree distribution



Hierarchical structure

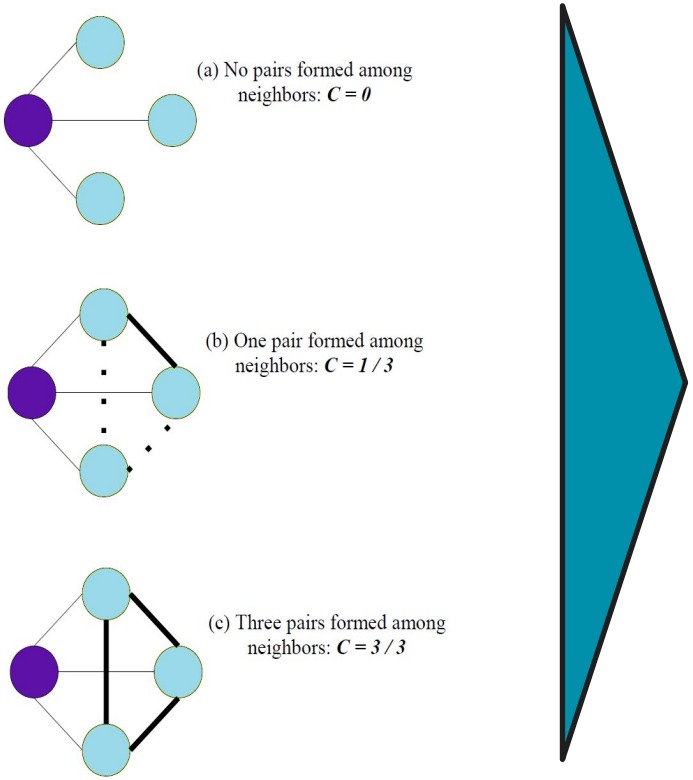
Gini coefficient of the in-degree distribution

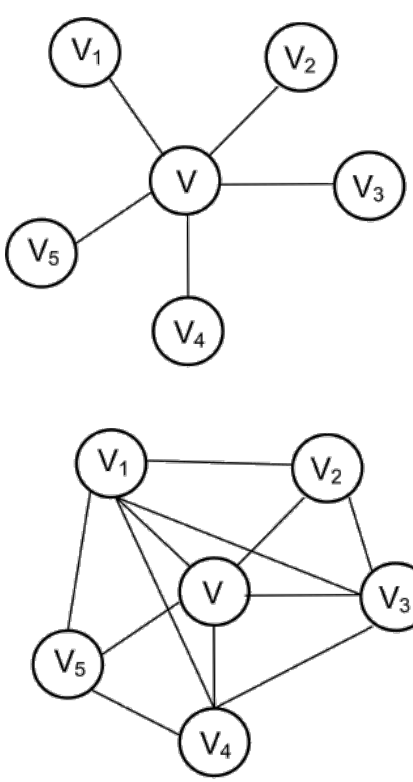


Cluster	G_{in}
BeC-p	0.995
Cs	0.964
ERC	0.954
CUP	0.953
CiU	0.893
PP	0.876
PSC	0.818
BeC-m	0.811

Small world phenomenon

Clustering coefficient



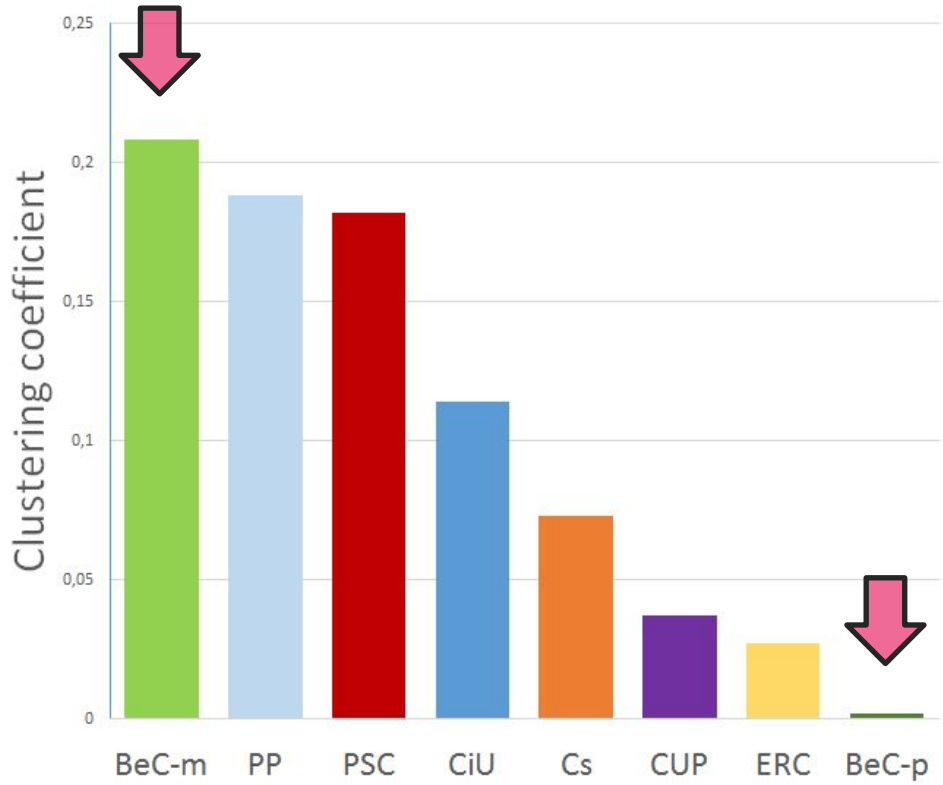


LOW

HIGH

Small world phenomenon

Clustering coefficient

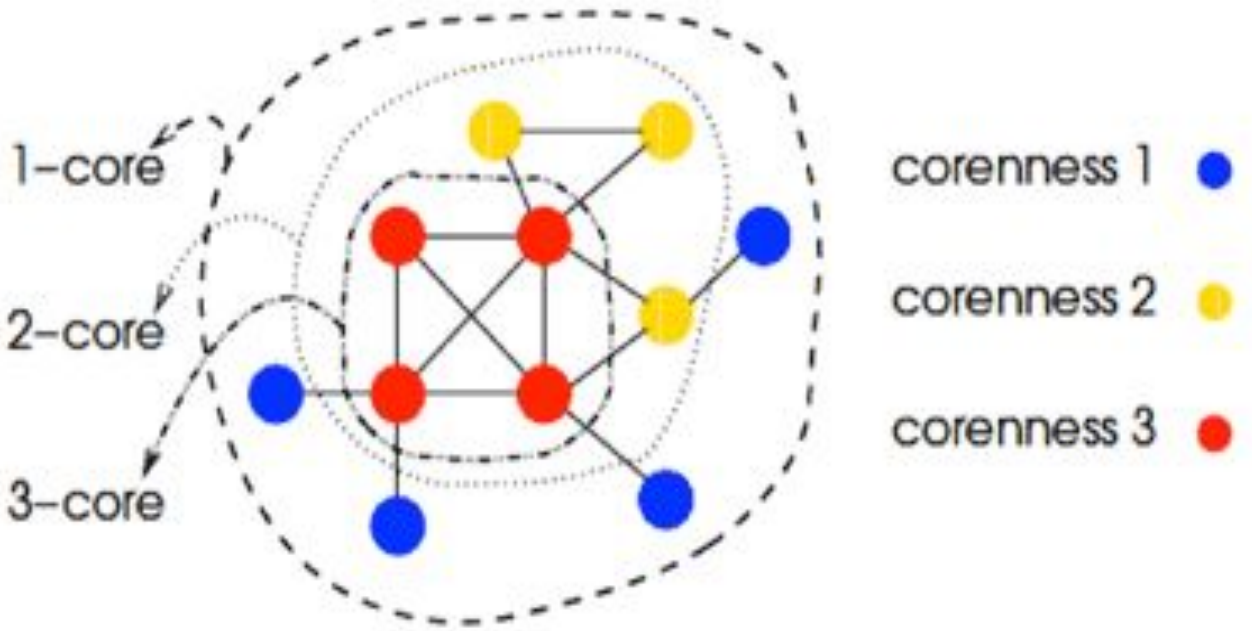


Number of nodes (N) and edges (E), clustering coefficient (Cl) and average path length (l) of each cluster.

Cluster	N	E	Cl	l
BeC-m	427	2 431	0.208	3.35
PP	301	1 163	0.188	2.73
PSC	211	810	0.182	2.29
CiU	337	1 003	0.114	4.66
Cs	352	832	0.073	2.57
CUP	635	1 422	0.037	2.57
ERC	866	1 899	0.027	5.43
BeC-p	1 844	2 427	0.002	2.48

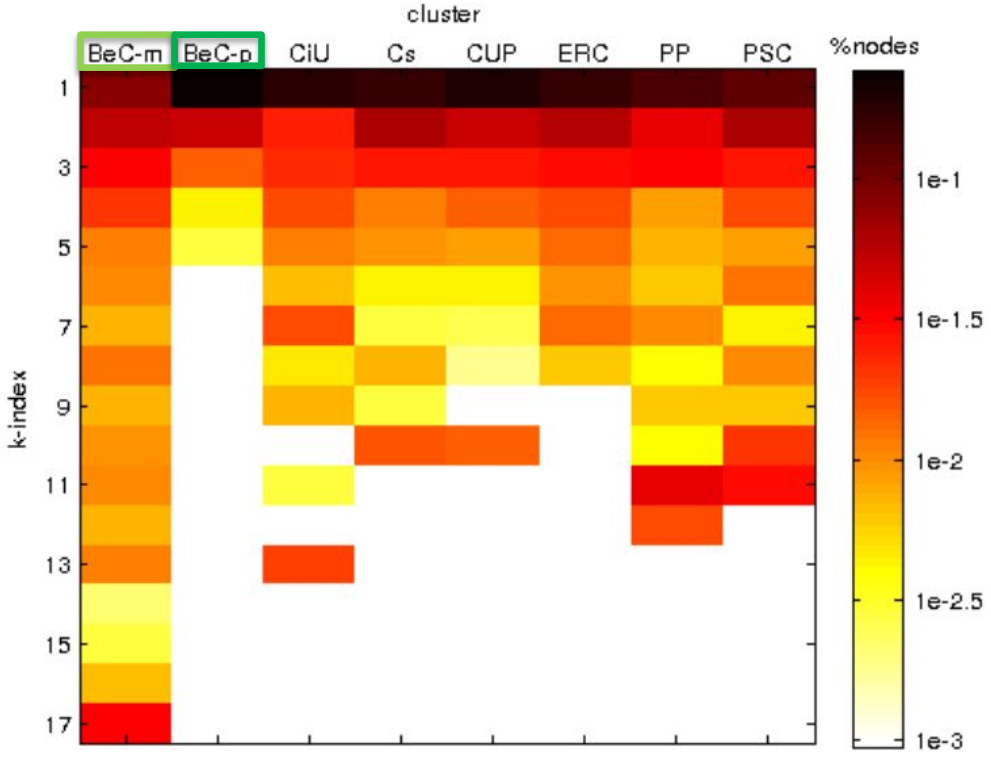
Coreness

K-core decomposition



Coreness

K-core decomposition



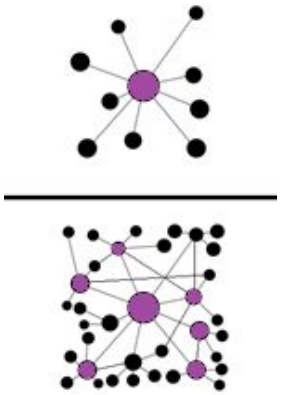
Maximal and average k-index (standard deviation in parentheses) for the intra-network of each cluster.

cluster	k_{max}	k_{avg}
BeC-m	17	5.90 (5.46)
PP	12	4.02 (3.99)
PSC	11	3.85 (3.55)
CiU	13	3.10 (3.44)
ERC	8	2.25 (1.85)
Cs	10	2.42 (2.42)
CUP	10	2.19 (2.22)
BeC-p	5	1.33 (0.71)

Conclusions

For Barcelona en Comú, two paradigms co-exist:

- A **centralized** and low resilient party cluster
- A **decentralized** and resilient movement cluster



Polarized scenario likes previous studies of election campaigns on Twitter

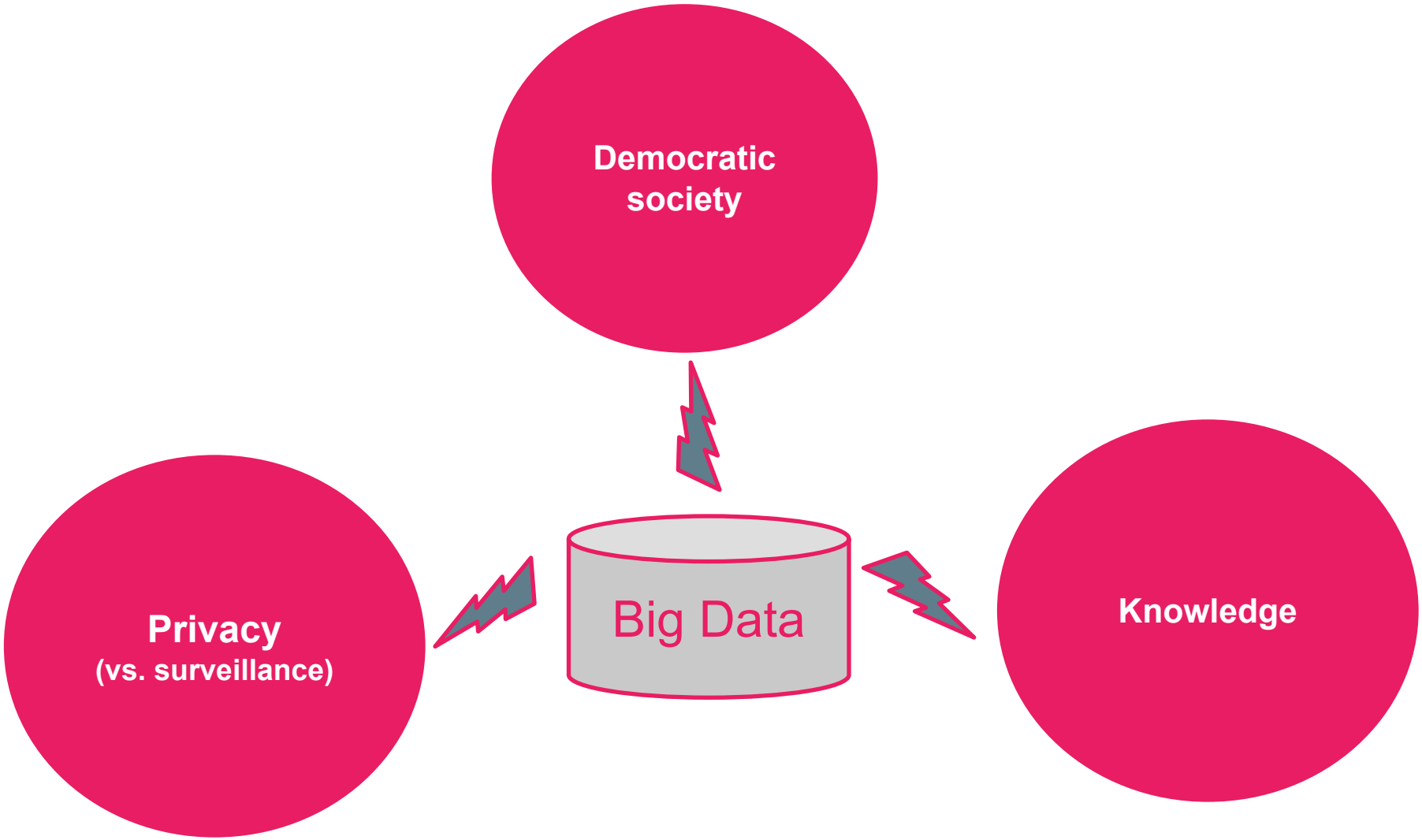
- Data preparation process accentuated the polarization effect

Media accounts build weak ties between clusters

- Public media became more plural than private media

Critical Questions for Big Data

Boyd, D., & Crawford, K. (2012).
Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon.
Information, communication & society, 15(5), 662-679.


Democratic society
Privacy
(vs. surveillance)
Big Data
Knowledge

Boyd & Crawford (2012)

———

“Big Data has emerged a system of knowledge that is already changing the objects of knowledge, while also having the power to inform how we understand human networks and community. ‘Change the instruments, and you will change the entire social theory that goes with them’, Latour (2009) reminds us.

(…)

Rather, it is a profound change at the levels of epistemology and ethics. Big Data reframes key questions about the constitution of knowledge, the processes of research, how we should engage with information, and the nature and the categorization of reality.”

Boyd & Crawford (2012)

- Big data changes the definition of knowledge
- Attributing great objectivity and precision to Big Data is misleading
- More data does not mean better data
- Without context, (Big) Data loses its meaning
- Its accessibility does not imply that its research is ethical
- Limited access to Big Data creates new digital divides

Boyd & Crawford (2012)

Does data speak for themselves?

NO! **Datasets and technology are biased**

Boyd & Crawford (2012)

“In reality, working with Big Data is still subjective, and what it quantifies does not necessarily have a closer claim on objective truth”

- There is a risk the Big Data in widening the division between "subjective" qualitative research and "objective" quantitative research.
- The process and analysis of Big Data contains many consecutive steps that are sometimes not recognized as subjective:
  - Data cleaning.
  - Methods of analysis and how they are applied.
  - Interpretation of the results.

Boyd & Crawford (2012)

“In reality, working with Big Data is still subjective, and what it quantifies does not necessarily have a closer claim on objective truth”

- What about the reliability of the datasets?
  - Errors in the datasets
- Data capture is generally very non-transparent
  - Biases and limitations of the data

Boyd & Crawford (2012)

Just because Big Data presents large amounts of data does not mean that the methodology is not relevant.

Understanding data is more important now than ever:

- Validity
- Reliability
- Does it fit the research questions?

Boyd & Crawford (2012)

Twitter is an excellent example of the limitations and biases of Big Data:

- It doesn't represent the whole population even though millions of people appear on the dataset
- There is no visibility of the dataset sampling
- Size is not equivalent to representativeness
- Restricted access to Twitter firehose, etc...

Boyd & Crawford (2012)

Data and digital methods do not have to be transferable from one context to another:

- The Facebook Graph may mean something on Facebook, but it's not a complete representation of people's real social network.
- The activity and intensity in the social network context may not have the same meaning in real life.

Big Data are not generic data about social interactions in general, they are specific to the given network.

Boyd & Crawford (2012)

Only social media firms have full access to data, researchers in general don't

Access to data is expensive → Inequalities for research

- Top universities improve their capabilities

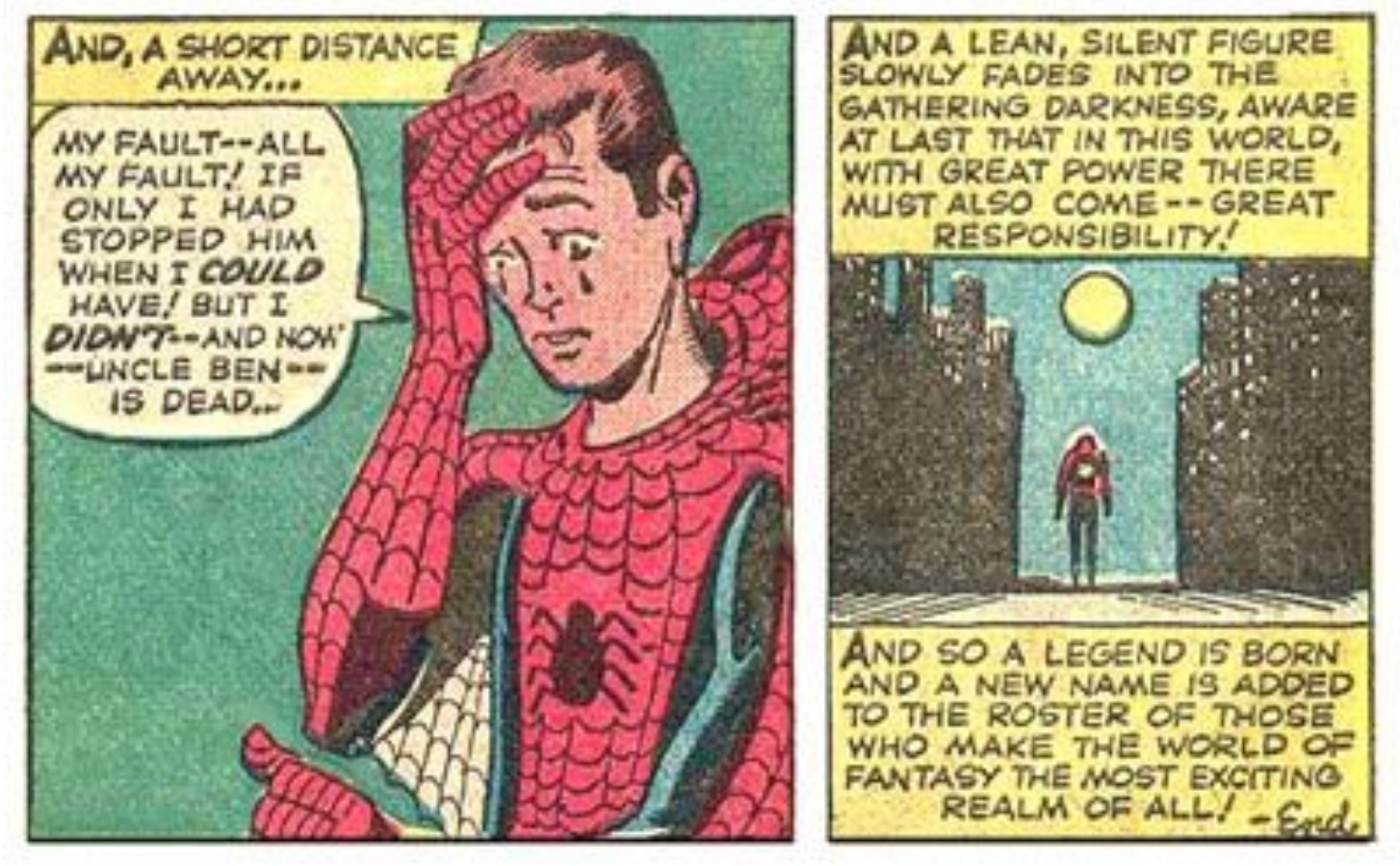
The skills required for access to data are limited to researchers with computer skills:

- Gender divide (among others)

Limited access creates a great bias in relation to the questions:

- Who decides the purposes of the Big Data?

Conclusion(s)



Homework

Watch the following documentary:

Orlowski, J. (2020)

The Social Dilemma

Netflix

https://www.netflix.com/es/title/81254224