

# Network analysis and visualization for social media

Andreas Kaltenbrunner

Social Media Research Group,  
Barcelona Media,  
Barcelona, Spain

School of advanced sciences of Luchon, July 3rd, 2014



## 1 Examples

## 2 Practical Session 1: Basics of Gephi

- Download <http://gephi.org/download/>
- Example network:

<http://gephi.org/datasets/LesMiserables.gexf>

## 3 Practical Session 2: Create and visualize your own networks

OR

Modeling the structure and evolution of online discussion cascades



# Part I: Examples for Network Analysis and Visualisation in Social Media



# Outline Part I

- 1 Political User interaction on Twitter
- 2 Political Affiliation on Wikipedia
- 3 Emotional styles on Wikipedia
- 4 Geographical distance and Friendship
- 5 Sister Cities
- 6 Links between biographies on Wikipedia



- 1 Political User interaction on Twitter
- 2 Political Affiliation on Wikipedia
- 3 Emotional styles on Wikipedia
- 4 Geographical distance and Friendship
- 5 Sister Cities
- 6 Links between biographies on Wikipedia



# Analysis of the Spanish General Elections of 2011

## Introduction

### Research Questions

- Do political parties interact on Twitter?
- Do political parties use Twitter to engage in conversations or as one-way flow broadcast medium?
- Are there differences between the parties?

### Dataset collected between Nov 4 and 24, 2011

- ~ 3 million tweets.
- ~ 380.000 users.

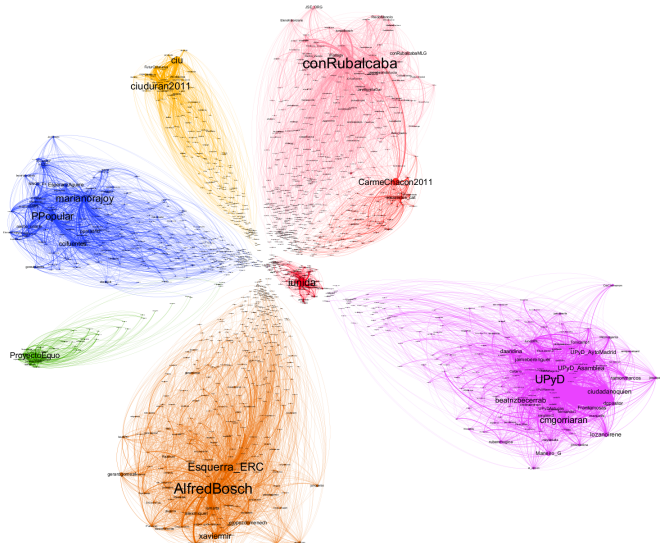
### Results published in



P. Aragón, K. Kappler, A. Kaltenbrunner, D. Laniado and Y. Volkovich.  
Communication Dynamics in Twitter During Political Campaigns: The Case of the 2011 Spanish National Election,  
*Policy & Internet*, 5 (2), 2013.

# Retweets

Users almost exclusively propagated contents from members of their own party



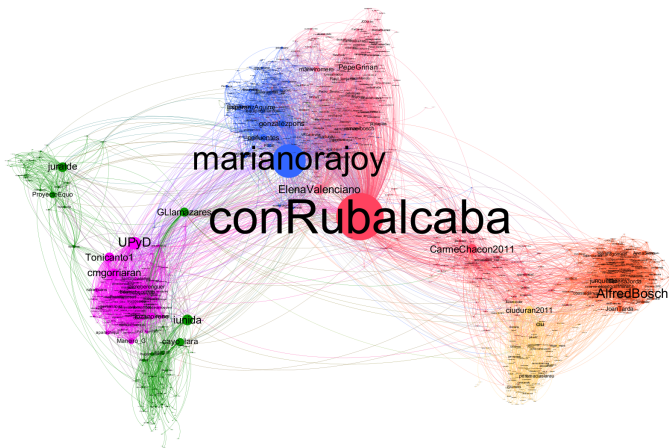
## Political parties

- PSOE
- PP
- EQUO
- IU
- ERC
- CiU
- UPyD



# Replies

The most intensive communication flows occur between members of the same party



- ### Political parties
- PSOE
  - PP
  - IU+EQUO
  - ERC
  - CiU
  - UPyD

Some amount of communication also among members of

● PP - PSOE

● IU - UPyD - EQUO

● ERC - CiU



## Conclusions

- Retweets: Balkanisation of Spain's (online) political sphere
- Replies: Inter-party communication happens but most of the interactions still occur within the parties.
- Political parties use Twitter as a one-way flow broadcast.
  - Low number of replies by candidate and party profiles
  - Low ratio between sent and received replies.
- New and minor parties tend to be more clustered and better connected  $\Rightarrow$  a more cohesive community.

## Future Research

- In-depth analysis of the topological patterns of party networks to characterise the different party apparatus (centralised, decentralised, or distributed).

# Outline

- 1 Political User interaction on Twitter
- 2 Political Affiliation on Wikipedia**
- 3 Emotional styles on Wikipedia
- 4 Geographical distance and Friendship
- 5 Sister Cities
- 6 Links between biographies on Wikipedia



## Does political polarisation also take place in Wikipedia?

- Obtain a deeper understanding of online interaction and collaboration among members of distinct political parties.

## Research questions

- Do political users in Wikipedia exhibit a preference for interacting with members of their same political party?
- Do we see a division in patterns of participation along party lines?

## Results published in



J. J. Neff, D. Laniado, K. E. Kappler, Y. Volkovich, P. Aragón & A. Kaltenbrunner.

*Jointly They Edit: Examining the Impact of Community Identification on Political Interaction in Wikipedia.*

PLoS ONE, vol. 8, no. 4, page e60584, 2013.



WIKIPEDIA  
The Free Encyclopedia

Main page  
Contents  
Featured content  
Current events  
Random article  
Donate to Wikipedia

Interaction  
Help  
About Wikipedia  
Community portal  
Recent changes  
Contact Wikipedia

Toolbox  
Print/export  
Languages  
Česky  
Español  
Français  
한국어  
Kiswahili  
Norsk (bokmål)  
Русский  
Svenska  
ไทย

Article [Discussion](#)

Read [Edit](#) [View history](#)

Search

[Log in / create account](#)

## Presidency of Barack Obama

From Wikipedia, the free encyclopedia

See also: *Timeline of the Presidency of Barack Obama*

The **Presidency of Barack Obama** began at noon EST on January 20, 2009 when he became the 44th President of the United States. Obama was a United States Senator from Illinois at the time of his victory over Arizona Senator John McCain in the 2008 presidential election. Obama became the first African-American president of the United States.

His policy decisions have addressed a global financial crisis and have included changes in tax policies, legislation to reform the United States health care industry, foreign policy initiatives and the phasing out of detention of prisoners at the Guantanamo Bay detention camp in Cuba. He attended the G-20 London summit and later visited U.S. troops in Iraq. On the tour of various European countries following the G-20 summit, he announced in Prague that he intended to negotiate substantial reduction in the world's nuclear arsenals, en-route to their eventual extinction. In October 2009, Obama was awarded the Nobel Peace Prize for "his extraordinary efforts to strengthen international diplomacy and cooperation between peoples."

### Contents [hide]

- Transition period
- Inauguration
- First 100 days
  - Expectations
  - Legislation and executive orders
- Approval ratings and opinion
  - 2009
    - Handling of the economy
  - 2010
- Major legislation
  - Legislation signed
    - 2009
    - 2010
    - 2011
- Personnel
  - Cabinet appointees
  - Notable non-Cabinet positions
  - Judicial nominees
    - Supreme Court
    - Other courts
- Policies
  - Economy
  - Ethics
    - Lobbying reform
    - Transparency
  - Foreign policy
    - Guantánamo Bay detention camp
    - Overseas Contingency Operation
  - Gun control
  - Science and technology
    - Cybersecurity
    - Environment
    - NASA

### Presidency of Barack Obama



**44th President of the United States**

**Incumbent**

**Assumed office**

January 20, 2009

**Vice President** Joe Biden

**Preceded by** George W. Bush

**Born** August 4, 1961 (age 49)  
Honolulu, Hawaii, United States<sup>[1]</sup>

**Birth name** Barack Hussein Obama II<sup>[1]</sup>

**Nationality** American

**Political party** Democratic

**Spouse(s)** Michelle Obama (m. 1992)

**Children** Malia Ann (b. 1998)  
Natasha (Sasha) (b. 2001)

**Residence** The White House

**Alma mater** Occidental College  
Columbia University

# Introduction

## Article talk pages



WIKIPEDIA  
The Free Encyclopedia

Main page  
Contents  
Featured content  
Current events  
Random article  
Donate to Wikipedia

Interaction  
Help  
About Wikipedia  
Community portal  
Recent changes  
Contact Wikipedia

Toolbox  
Print/export

Log in / create account

Article Discussion

Read Edit New section View history Search

## Talk:Presidency of Barack Obama

From Wikipedia, the free encyclopedia

[Skip to table of contents](#)


**This is the talk page for discussing improvements to the Presidency of Barack Obama article.**


<ul style="list-style-type: none"><li><b>This is not a forum for general discussion of the article's subject.</b></li><li><b>Put new text under old text.</b> <a href="#">Click here to start a new topic.</a></li><li><b>Please sign and date your posts</b> by typing four tildes (----).</li><li><b>New to Wikipedia?</b> Welcome! Ask questions, get answers.</li></ul>	<ul style="list-style-type: none"><li>Be polite</li><li>Assume good faith</li><li>Avoid personal attacks</li><li>Be welcoming</li></ul>	<b>Article policies</b> <ul style="list-style-type: none"><li>No original research</li><li>Neutral point of view</li><li>Verifiability</li></ul>
---	---	--

**Archives:** 1, 2, 3, 4

**This article has been placed on article probation.** Editors making disruptive edits may be blocked temporarily from editing the encyclopedia, or subject to other administrative remedies, according to standards that may be higher than elsewhere on Wikipedia. Please see [Talk:Barack Obama/Article probation](#) for full information and to review the decision.

**Administrators:** when sanctioning an editor for disruption to an article under probation, please be sure to record the action in the appropriate log. The log is linked [here](#), under "decision and log" on the sanction's row in the table.

 This article is within the scope of **WikiProject Barack Obama**, a collaborative effort to improve the coverage of Barack Obama on Wikipedia. If you would like to participate, please visit the project page, where you can join the [discussion](#) and see a list of open tasks.



**B** This article has been rated as **B-Class** on the quality scale.

**Top** This article has been rated as **Top-importance** on the importance scale.

### Contents [hide]

- 2010 election results
- WikiLeaks release of secret documents
- 2010 elections
- Filibuster, etc.
- New source - Terrible, Horrible, etc.
- job approval ratings

## 2010 election results

I feel we should add something on here about the Republican victories in the 2010 elections and how that will be a turning point for Obama similar to the Presidency of Bill Clinton article. Where should we put that? [Politics2012](#) (talk) 03:39, 09 November 2010 (UTC)

How will it be a "turning point"? Do you have a *crystal ball* or something? -- [Scjessey](#) (talk) 21:53, 9 November 2010 (UTC)

I meant that we should put somewhere in here something about Republican victories in general. It says something in the Bill Clinton Presidency article about the 94 Republican victories and in the George W. Bush presidency article about the Democratic victories in the 2006 elections. We need to put something in here about the 2010 Republican victories. But where? [Politics2012](#) (talk) 03:39, 09 November 2010 (UTC)

I disagree. The information in those other articles has been written with the benefit of an historical perspective, but we do not have that luxury here. We must wait and see what sources say about how Republican

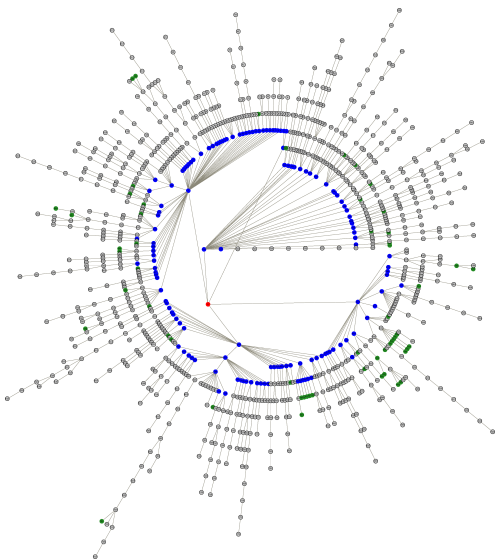


Archives  
1, 2, 3, 4

[edit]

# Example Structure

Discussion tree for article “Presidency of Barack Obama”



- red → root (the article)
- blue → structural nodes
- green → anonymous comments
- grey → registered comments

## More details in:



D. Laniado, R. Tasso, Y. Volkovich,  
and A. Kaltenbrunner.

When the Wikipedians talk: Network  
and tree structure of Wikipedia  
discussion pages.

*In Proc. of ICWSM, 2011.*

# Interactions of partisan users on article talk pages

User-boxes  $\Rightarrow$  Party assign.

## Democrats

**dem**

This user supports the **U.S. Democratic Party.**

## Republicans

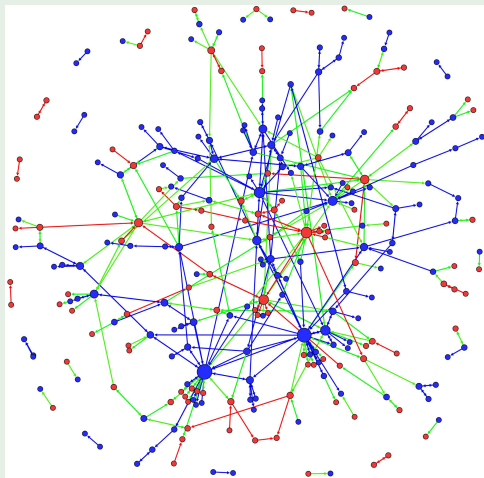


This user supports the **Republican Party of the United States.**

## Cross-party interactions

- Shuffle test indicates neutral mixing.
- $\Rightarrow$  no stat. significant preference for neither inter- nor intra-party interaction.

## Interaction Network



Democrats vs. Republicans

## Motivation

- Measures if there exists a preference for relations between users of the same or different characteristics.
- Possible characteristics:
  - Number of relations
  - Sex
  - Age
  - Race
  - Weight
  - Mother tongue
  - ...
- Examples can be found in [Newman 2003].





# Method: Calculate mixing coefficient with reshuffling I

Data: Pairs of users interacting broken by party.

article discussions	Democrats	Republicans
Democrats	193	94
Republicans	86	57
user wall	Democrats	Republicans
Democrats	395	243
Republicans	187	172

Definition: mixing coefficient

$$r = \frac{\text{Tr}A - \|A^2\|}{1 - \|A^2\|}$$

where  $A$  is a normalised matrix with elements  $a_{ij}$  and  $\|A^2\|$  is the sum over all  $a_{ij}^2$

# Calculate mixing coefficient with reshuffling II

Mixing coefficient  $r$

## Interpretation

$r > 0$ : assortative mixing

- There exists a preference for relations between similar users.
- Users with the same characteristics relate preferentially among themselves and vice versa.

$r \approx 0$ : neutral mixing

- There is no preference in the relations.

$r < 0$ : disassortative mixing

- There exists a preference for relations among users with different characteristics.
- For example between users with the opposite ideological views.

# Calculate mixing coefficient with reshuffling III

## To avoid bias due to network topology

- E. g. one group of users being more active than the other

## Compare with $r_{rand}$ in reshuffled networks

- keep the users fixed,
  - same party affiliations
  - same numbers of in-coming and out-going links
- randomise the links between them
- generate a sample of 100 networks
- computed the average mixing coefficient  $\hat{r}_{rand}$  of these networks and their standard deviation  $\sigma_{rand}$ .
- calculate Z-score

$$\text{Z-score} = (r - \hat{r}_{rand}) / \sigma_{rand}$$

# Calculate mixing coefficient with reshuffling IV

## Interpretation Z-score

- High positive values of Z indicate assortative mixing
- High negative values indicate dissortative mixing.
- Low absolute values ( $|Z| < 2$ ) correspond to neutral mixing, i.e. no statistically significant preferences [Foster 2010].

## Results

talk page	$r$	$\hat{r}_{rand}$	$\sigma_{rand}$	Z-score	significant?
article	0.070	0.0028	0.0505	1.33	no
user	0.095	-0.0053	0.0301	3.33	yes

## Conclusions

- Wikipedian identity seems to predominate over party identity in article discussions.

# Outline

- 1 Political User interaction on Twitter
- 2 Political Affiliation on Wikipedia
- 3 Emotional styles on Wikipedia**
- 4 Geographical distance and Friendship
- 5 Sister Cities
- 6 Links between biographies on Wikipedia



# Analysis of emotions expressed in talk pages

## Introduction

### Goal:

Study the emotional dimension in a large peer production community

Research questions: How are the emotional styles of editors ...

- 1 affected by their **level of experience**?
- 2 affected by their **gender** and the **topics** they choose to work on?
- 3 affected by interacting with others (**emotional congruence**)?
- 4 related to those of the editors they interact more frequently with (**emotional homophily**)?

Results are partly published in



Laniado, D., Castillo, C., Kaltenbrunner, A., and Fuster Morell, M. F. (2012)  
Emotions and dialogue in a peer-production community: the case of Wikipedia.  
*8th International Symposium on Wikis and Open Collaboration, WikiSym'12*

# User gender labelling

- $\approx 12\,000$  users wrote  $\geq 100$  comments in articles talk pages
- Gender identified through Wikipedia API for  $\approx 2\,000$  of them
- Out of the remaining ones, a sample of 1 385 users for manual labelling through crowd-sourcing (Crowdflower)

The screenshot shows the Wikipedia user page for 'User:Shell Kinney'. At the top, there is a navigation bar with 'User page' and 'Talk' tabs, and a search box. Below this, the user's name 'User:Shell Kinney' is displayed, followed by the text 'From Wikipedia, the free encyclopedia'. A large black banner with the word 'RETIRED' in white capital letters is centered on the page. Below the banner, a message states 'This user is no longer active on Wikipedia.' and 'OFFLINE'. A row of icons for various user actions is visible, including 'Talk page', 'Contact', 'Email', 'Adoptees', 'Archives', 'Articles', 'Watching', 'Awards', 'Logs', 'Sandbox', and 'Userspace'. Below this, a grey box contains the text 'Wait - where did my life go?'. The main content area is divided into two columns. The left column has a green header 'Oh hai! My admnim skills, let me show you them' and contains two images: a grumpy-looking cat with the text 'I IZ SERIUS ADMNIM THIZ IZ SERIUS BIZNIS' and a white pig looking at a laptop with the text 'NEEDS MOAR DRAMA'. The right column has a green header 'Quick info' and contains a grid of six colored boxes with user statistics: 'This user is a female', 'This user assumes good faith.', 'This user is a Placem...', 'My besting alternate account is User:Shell Kinney (watch)', 'This user practices an eclectic form of Buddhism / Earth mother mysticism.', 'This user is owned by lots and lots of cats.', 'This user has over 32,000 edits.', and 'This user has been on Wikipedia for 6 years, 7 months and 13 days.' The left sidebar contains the Wikipedia logo and a list of navigation links.

# User gender labelling

## Manual labelling

- Gender could be identified only for  $\approx 50\%$  of users:
  - real name or username (50% of those identified)
  - implicitly stated gender (27% of females, 20% of males)
  - pronoun (15% of females, 10% of males)
  - other indicators: userboxes, pictures, links to personal blogs...

	Non-admins	Admins	Total
Males	1 087	1 526	2 613
Females	68	97	165
Unknown	6 850	2 603	9 453
Total	8 005	4 226	12 231

**Table:** Users with  $\geq 100$  comments by gender and administrator status.

- Category “unknown” includes:
  - 8 708 users that were not included in the crowd-sourced task
  - 745 users whose gender could not be identified by evaluators

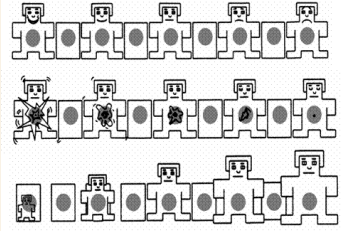




# Measuring the Emotional Content of Discussions

## Affective norms for English words (ANEW)

Rates a list of 1060 frequent words on a 9 point scale in three dimensions:

- Valence
  - Arousal
  - Dominance
- 
- Compare users per word frequency-weighted averages.



Bradley and Lang. (1999).

Affective norms for English words (ANEW) Technical report C-1.

*The Center for Research in Psychophysiology, University of Florida, FL.*

## Linguistic Inquiry and Word Count (LIWC)

- Discrete measures of emotions (anger, anxiety and sadness)
- Two scores for basic emotion (compared with ANEW valence)
  - positive valence and
  - negative valence
- by counting the proportion of positive / negative words in a comment
- ANEW assigns emotions scores to each word from the lexicon.



Pennebaker J, Chung C, Ireland M, Gonzales A, Booth R (2010).

The development and psychometric properties of LIWC2007. Austin, TX.



## SentiStrength

- Based on LIWC and developed for short web texts
- Accounts for modes of textual expression specific to the online environment, e.g. emoticons and abbreviations.
- Provides a positive and a negative score for emotional valence.
- Emotion score is calculated at the sentence level (number of positive and negative words).
- Summarised at the comment level as strongest positive and negative emotion expressed in a comment.
- Final scores are averages over comments in a given category.

 Thelwall M, Buckley K, Paltoglou G, Cai D, Kappas A (2010)

Sentiment strength detection in short informal text.

*Journal of the American Society for Information Science and Technology* 61: 2544 – 2558.



# Measuring the Emotional Content of Discussions

Example for the results of different emotional lexica

**Table:** Example messages with their corresponding Valence, Arousal, and Dominance (ANEW) or positive & negative scores (LIWC, **SentiStrength**).

	ANEW			LIWC		SentiSt.	
	V	A	D	+	-	+	-
Sounds like a <b>good challenge</b> - to be proven or disproven. I'm <b>happy</b> if it can be shown to go further using closed cubic polynomial solutions. The <b>nice</b> thing about these are that they are <b>pretty easy</b> to test numerically . . . -in "Exact trigonometric constants"	7.4	5.3	6.2	15	0	3	-2
Seems you have not yet seen female <b>lover</b> after having <b>sex</b> who do not <b>wish</b> to have <b>sex</b> with the same <b>lover</b> any more :) Once you've seen it, you understand very <b>well</b> what <b>war</b> of Venus means compared to <b>war</b> of Mars. -in "House (astrology)"	5.5	7.0	5.2	6.8	4.5	4	-3
What about the whirlie hazing, the alcohol <b>abuse</b> , the <b>emotional poverty</b> , the <b>suicide</b> in 1995/6, the biotech plans which were stopped by pitzer <b>protests</b> . . . -in "Harvey Mudd College"	1.6	5.8	3.5	4	8	1	-4



# Emotions, Status and Gender

Similar results with different Lexica

## Emotions and Status

- Admins express, on average, more positive emotion ( $p < 0.001$ ).
- Admins also express less negative emotion ( $p < 0.001$ ).
- Non-admins express more affect, in particular, more anxiety, anger and sadness (all with  $p < 0.001$ ) compared to admins.

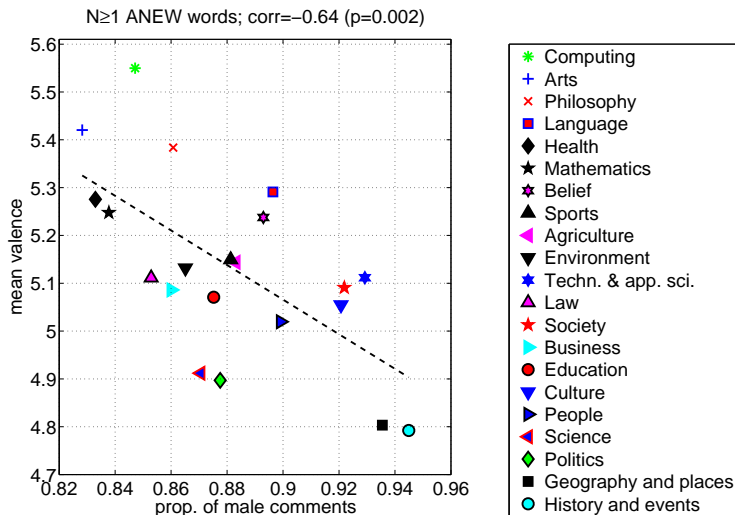
## Emotions, Status and Gender

- Significant difference between male admins and non-admins ( $p < 0.001$ )
- No significant difference between female admins and non-admins.





# Topics, emotions and gender



**Figure:** Mean valence for discussions of articles in different topic categories, vs the proportion of comments written by male editors

# Emotional congruence

Replies are more positive

## On average, editors tend to reply with:

- higher valence:  $+0.05$  ( $p < 0.01$ )
  - higher dominance:  $+0.04$  ( $p < 0.01$ )
  - no statistically significant differences for arousal
- 
- Users tend to be more positive and dominant when replying, but without recurring to words evoking stronger sentiments.

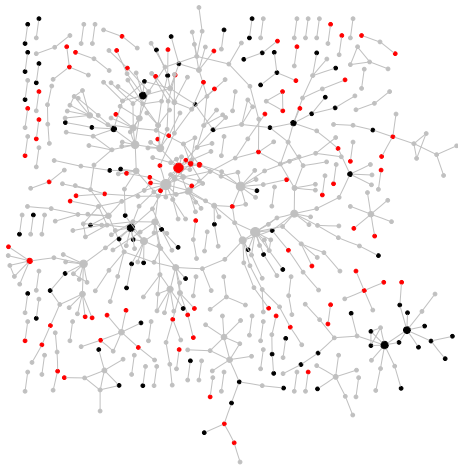




# Emotional homophily

Mixing patterns: do users interact preferentially with similar users?

- *Assortative by emotional style*: users interact more with others expressing similar emotions.



- edges connecting users who have exchanged at least 10 replies
- **red nodes** → 15% users expressing higher valence in article discussions
- **black nodes** → 15% users expressing lower valence in article discussions
- size → proportional to the number of connections



# Homophily: detailed results

Normalised		$r$	$r_{rand}$	$\sigma_{rand}$	$Z$
valence	(sent)	0.0269	-0.0003	0.0011	<b>23.8</b>
	(received)	0.0109	-0.0004	0.0010	<b>10.8</b>
arousal	(sent)	0.0253	-0.0004	0.0009	<b>28.2</b>
	(received)	0.0187	0.0013	0.0012	<b>14.8</b>
dominance	(sent)	0.0380	-0.0001	0.0015	<b>26.2</b>
	(received)	0.0121	9.8e-08	0.0011	<b>10.8</b>

	$r$	$r_{rand}$	$\sigma_{rand}$	$Z$
gender	0.0443	-0.0008	0.0059	<b>7.63</b>
#comments written	-0.0177	-0.0014	0.0017	<b>-9.51</b>
#replies received	-0.0060	-0.0013	0.0014	<b>-3.50</b>
#replied users	-0.0340	-0.0023	0.0020	<b>-16.23</b>
#replying users	-0.0237	-0.0014	0.0015	<b>-14.35</b>
#discussed articles	-0.0009	-0.0011	0.0014	0.12

# Outline

- 1 Political User interaction on Twitter
- 2 Political Affiliation on Wikipedia
- 3 Emotional styles on Wikipedia
- 4 Geographical distance and Friendship**
- 5 Sister Cities
- 6 Links between biographies on Wikipedia



# Motivation

## Distance and friendship

### online tools and long-distance travel $\Rightarrow$ *death of distance*?

- individuals try to minimise the efforts to maintain a friendship by interacting more with their spatial neighbours.
- probability of a social interaction quickly decays as an inverse power of the relative geographic distance [Stewart 1941].
- probability of connections between two individuals on online social networking services still decreases with their geographic distance [Backstrom 2010, Liben-Nowell 2005].

## Results published in



A. Kaltenbrunner, S. Scellato, Y. Volkovich, D. Laniado, D. Currie, E. J. Jutemar & C. Mascolo.  
*Far from the eyes, close on the Web: impact of geographic distance on online social interactions.*  
In Proceedings of ACM SIGCOMM Workshop on Online Social Networks (WOSN '12). ACM, 2012.

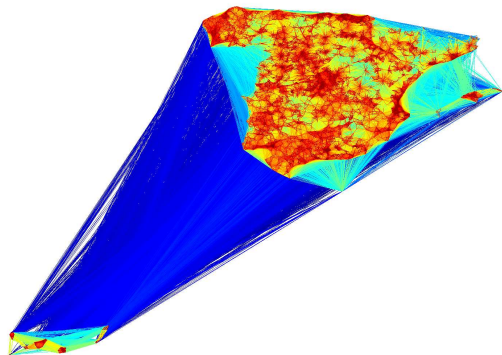


Y. Volkovich, S. Scellato, D. Laniado, C. Mascolo & A. Kaltenbrunner.  
*The length of bridge ties: structural and geographic properties of online social interactions.*  
In ICWSM-12 - 6th International AAAI Conference on Weblogs and Social Media. The AAAI Press, 2012.

# Dataset from Tuenti

“Spanish Facebook”, a Spain-based social networking website

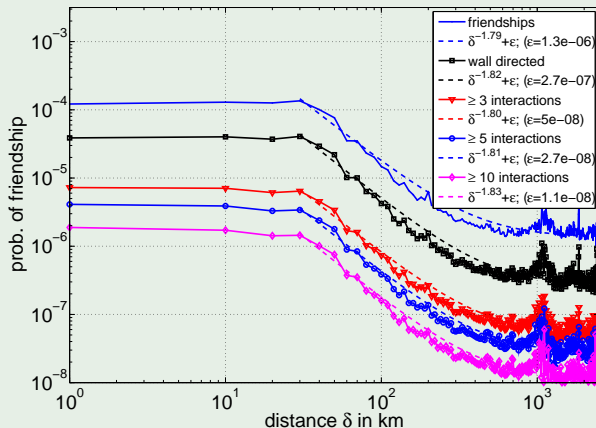
- ~10 million users
- 1174 million friendship links
- ~500 directed messages exchanges during 3 months;



# Geographic properties

The effect of distance on friendship

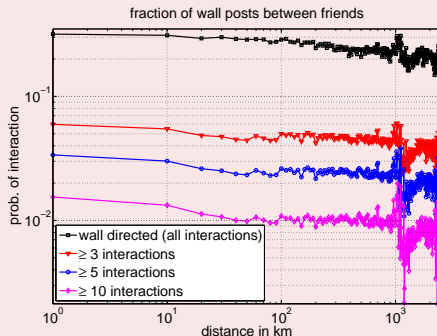
## Probability of connection as function of geographic distance



# Interaction Analysis

## Interactions and distance

### Probability of message exchange between friends



- high-intensity communication takes place on social connections regardless of their geographic distance

## The effect of geographic distance on online social interactions

- Spatial proximity greatly affects how users establish their connections on online social platforms.
- Social interactions are only weakly affected by distance.
- Geography affects **whom** we interact with, however it does not influence **how much** we interact.

## Applications

- link prediction,
- tie strength modelling,
- user profiling.





# Outline

- 1 Political User interaction on Twitter
- 2 Political Affiliation on Wikipedia
- 3 Emotional styles on Wikipedia
- 4 Geographical distance and Friendship
- 5 Sister Cities**
- 6 Links between biographies on Wikipedia



# Analysis of institutional (sister city) relations



# Introduction

## Analysis of institutional (sister city) relations

### Sister cities

- Institutional partnership between two cities or towns with the aim of cultural and economical exchange.
- These relations had never been analysed before.

### We want to understand ...

- social
- geographical
- economic mechanisms

of city pairings.



### Results published in



Andreas Kaltenbrunner, Pablo Aragón, David Laniado & Yana Volkovich.

*Not All Paths Lead to Rome: Analysing the Network of Sister Cities.*

In *Self-Organizing Systems, Lecture Notes in Computer Science*, vol. 8221, Springer, 2014.

# Example for Wikipedia article used for data extraction

## List of twin towns and sister cities in Spain

From Wikipedia, the free encyclopedia

This is a list of places in **Spain** having standing links to local communities in other countries. In most cases, the association, especially when formalised by local government, is known as "**town twinning**" (though other terms, such as "partner towns" or "sister cities" are sometimes used instead), and while most of the places included are towns, the list also comprises villages, cities, districts, counties, etc. with similar links.

### A Coruña

[\[edit\]](#)

- Brest, Brittany, France
- Recife, Pernambuco, Brazil

### Antequera

[\[edit\]](#)

- Agde, Hérault, France

### Almansa

[\[edit\]](#)

- Lymington, United Kingdom

### Alcalá de Henares

[\[edit\]](#)

- Alba Iulia, Romania
- Peterborough, United Kingdom

### Altea

[\[edit\]](#)

**Altea** is a founding member of the **Douzelage**, a **town twinning** association of 23 towns across the European Union. This active town twinning began in 1991 and there are regular events, such as a produce market from each of the other countries and festivals.<sup>[1][2]</sup>

- Bad Kötzing, Germany<sup>[2]</sup>
- Bellagio, Italy<sup>[2]</sup>
- Bundoran, Republic of Ireland<sup>[2]</sup>
- Chojna, Poland<sup>[2]</sup>
- Granville, France<sup>[2]</sup>
- Holstebro, Denmark<sup>[2]</sup>
- Houffalize, Belgium<sup>[2]</sup>
- Judenburg, Austria<sup>[2]</sup>
- Karkkila, Finland<sup>[2]</sup>
- Kőszeg, Hungary<sup>[2]</sup>
- Marsaskala, Malta<sup>[2]</sup>
- Meerssen, the Netherlands<sup>[2]</sup>
- Niederanven, Luxembourg<sup>[2]</sup>
- Oxelösund, Sweden<sup>[2]</sup>
- Prienai, Lithuania<sup>[2]</sup>
- Preveza, Greece<sup>[2]</sup>
- Sesimbra, Portugal<sup>[2]</sup>
- Sherborne, United Kingdom<sup>[2]</sup>
- Sigulda, Latvia<sup>[2]</sup>
- Sušice, Czech Republic<sup>[2]</sup>
- Türi, Estonia<sup>[2]</sup>
- Zvolen, Slovakia<sup>[2]</sup>

### Ames

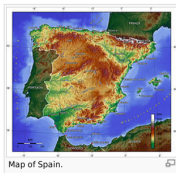
[\[edit\]](#)

Sahara Desert

### Archena

[\[edit\]](#)

- Chesham, United Kingdom (1995)<sup>[3]</sup>



# Dataset extracted from the English Wikipedia

## Data extraction process

- automated parser and a manual cleaning process.
- Google Maps API to geo-locate cities.

## Size of the dataset

network	$N$	$K$	$\langle C \rangle$	% GC	$\langle d \rangle$
city network	11 618	15 225	0.11	61.35%	6.74
country network	207	2933	0.43	100%	2.12

## Disclaimer

- No central register.
- User generated data (only 30% of reciprocal connections).
- No guarantee that the dataset is complete.

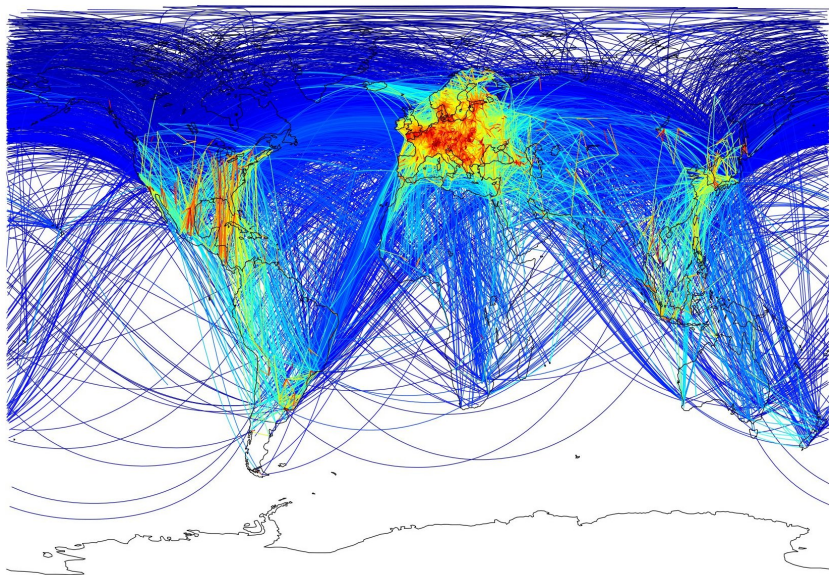
# Top 20 cities and countries ranked by degree

rank by betweenness centrality in parenthesis

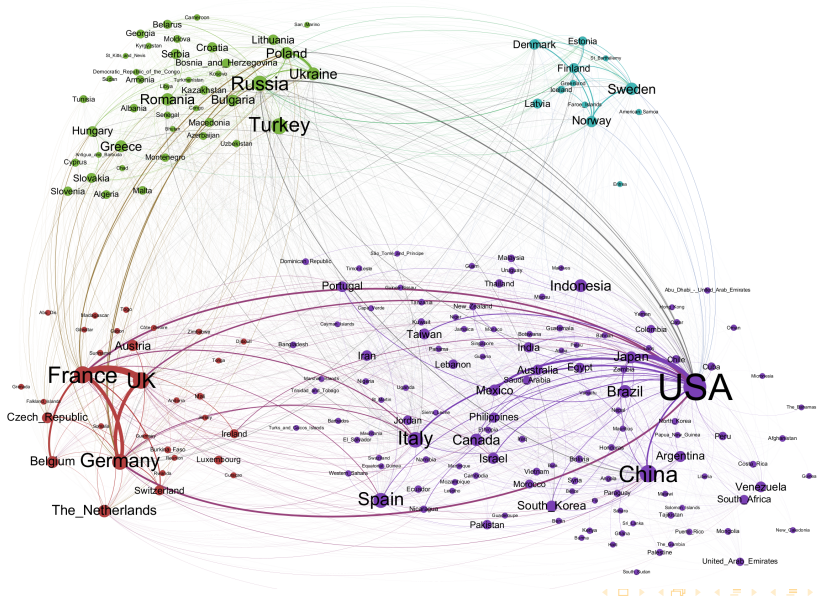
Nº	city	deg.	betw.	country	w. deg.	betw.
1	Saint Petersburg	78	(1)	USA	4520	(1)
2	Shanghai	75	(4)	France	3313	(3)
3	Istanbul	69	(12)	Germany	2778	(6)
4	Kiev	63	(5)	UK	2318	(2)
5	Caracas	59	(23)	Russia	1487	(9)
6	Buenos Aires	58	(36)	Poland	1144	(33)
7	Beijing	57	(124)	Japan	1131	(20)
8	São Paulo	55	(24)	Italy	1126	(7)
9	Suzhou	54	(6)	China	1076	(4)
10	Taipei	53	(20)	Ukraine	946	(27)
11	Izmir	52	(3)	Sweden	684	(14)
12	Bethlehem	50	(2)	Norway	608	(22)
13	Moscow	49	(16)	Spain	587	(11)
14	Odessa	46	(8)	Finland	584	(35)
15	Malchow	46	(17)	Brazil	523	(13)
16	Guadalajara	44	(9)	Mexico	492	(21)
17	Vilnius	44	(14)	Canada	476	(28)
18	Rio de Janeiro	44	(29)	Romania	472	(32)
19	Madrid	40	(203)	Belgium	464	(23)
20	Barcelona	39	(60)	the Netherlands	461	(16)



# Sister city relations



# Clustering of relations aggregated by country





## Method

- Compare sister city network and 100 randomised equivalents.
- Calculate assortativity measure based on the Z-score

## Degree assortativity by city

- Cities with many connections tend to be connected with cities with many connections and vice-versa.

## Relations are assortative by country

- Gross Domestic Product per capita
- Human Development Index
- Political Stability Index



## Details

property	$r$	$r_{rand}$	$\sigma_{rand}$	$Z$
city degree	0.3407	-0.0037	0.0076	<b>45.52</b>
Gross Domestic Product (GDP) <sup>a</sup>	0.0126	-0.0005	0.0087	1.51
GDP per capita <sup>b</sup>	0.0777	0.0005	0.0078	<b>9.86</b>
Human Development Index (HDI) <sup>c</sup>	0.0630	-0.0004	0.0075	<b>8.46</b>
Political Stability Index <sup>d</sup>	0.0626	0.0004	0.0090	<b>6.94</b>

<sup>a</sup>Source

[http://en.wikipedia.org/wiki/List\\_of\\_countries\\_by\\_GDP\\_\(nominal\)](http://en.wikipedia.org/wiki/List_of_countries_by_GDP_(nominal))

<sup>b</sup>Source: [http://en.wikipedia.org/wiki/List\\_of\\_countries\\_by\\_GDP\\_\(nominal\)\\_per\\_capita](http://en.wikipedia.org/wiki/List_of_countries_by_GDP_(nominal)_per_capita)

<sup>c</sup>Source: [http://en.wikipedia.org/wiki/List\\_of\\_countries\\_by\\_Human\\_Development\\_Index](http://en.wikipedia.org/wiki/List_of_countries_by_Human_Development_Index)

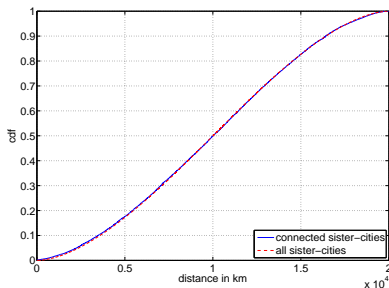
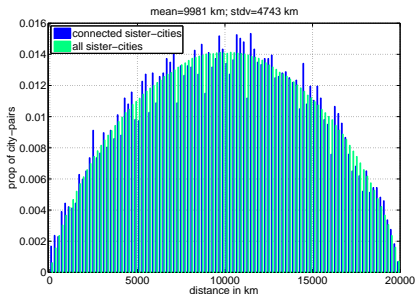
<sup>d</sup>Source: [http://viewswire.eiu.com/site\\_info.asp?info\\_name=social\\_unrest\\_table](http://viewswire.eiu.com/site_info.asp?info_name=social_unrest_table)



# Distances between sister cities

## Comparison of distances between two pairs of ...

- **connected** sister cities
- **random** (not necessarily connected) cities



## First evidence for the *Death of Distance*

- Nearly no differences between the two distributions.

## Conclusions

- Assortative mixing with respect to degree, economic and political country indexes.
- Sister city relationships reflect country predilections in and between cultural clusters.
- Geographic distance between cities does not influence city pairing.

## Future work

- Combined analysis with networks of air traffic or good exchange.
- Analysis of network evolution (needs other data-sources).



# Outline

- 1 Political User interaction on Twitter
- 2 Political Affiliation on Wikipedia
- 3 Emotional styles on Wikipedia
- 4 Geographical distance and Friendship
- 5 Sister Cities
- 6 Links between biographies on Wikipedia**



## Wikipedia as global collective memory place allows ...

- to extract from biographies how social links are recorded ...
- 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?

## Results published in

 P. Aragón, A. Kaltenbrunner, D. Laniado & Y. Volkovich.

*Biographical Social Networks on Wikipedia - A cross-cultural study of links that made history.*

In Proc. of the 8th Int. Symp. on Wikis and Open Collaboration (WikiSym'12). ACM, 2012.

# Data extraction

## Building biographical networks for 15 language editions of Wikipedia

- 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
- Manage alternative titles of articles: track redirects
- Data collected through Wikipedia APIs between September 8th and 13th, 2011



# Most central persons in the English Wikipedia

sorted by in-degree. Ranks for out-degree, betweenness and PageRank in parenthesis

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)





# Most central persons in different language Wikipedias

Top 5 most central persons for each language by betweenness

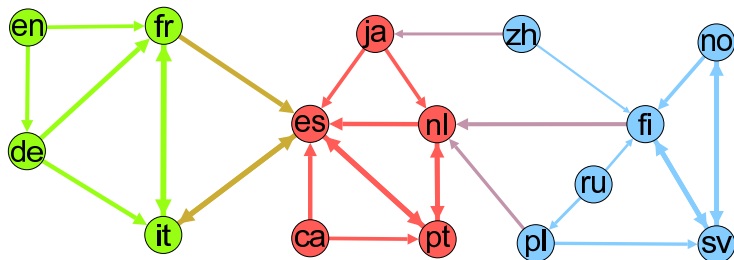
lang	#1	#2	#3	#4	#5
en	George W. Bush	Ronald Reagan	Adolf Hitler	Bill Clinton	John F. Kennedy
de	Adolf Hitler	George W. Bush	Martin Luther King, Jr	Barack Obama	Frank Sinatra
fr	Adolf Hitler	George W. Bush	William Shakespeare	Barack Obama	Jacques Chirac
it	Frank Sinatra	George W. Bush	Pope John Paul 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

Most are known to be (or have been) highly influential

- 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 peculiarities.

# Languages similarity network

Every language links to the two most similar ones according to Jaccard coefficient



## Definition of Jaccard coefficient $J$

- Given the set of links  $A$  and  $B$  of two networks

$$J = \frac{|A \cap B|}{|A \cup B|}$$

- $J$  is the ratio between the number of links present in both networks (their intersection) and the number of links existing in their union.



# Conclusions and future work

## 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.

## Future work

- Application of the methodology to generate subnetworks of other kinds of article categories
- Consider all biographies for each language.
- Analyse links missing only in a few language Wikipedias.

# Questions?



# Bibliography I



P. Aragón, A. Kaltenbrunner, D. Laniado & Y. Volkovich.

*Biographical Social Networks on Wikipedia - A cross-cultural study of links that made history.*

In Proceedings of the 8th International Symposium on Wikis and Open Collaboration (WikiSym'12). ACM, 2012.



P. Aragón, K. Kappler, D. Laniado A. Kaltenbrunner & Y. Volkovich.

*Communication Dynamics in Twitter During Political Campaigns: The Case of the 2011 Spanish National Election.*

Policy & Internet, vol. 5, no. 2, 2013.

in press.



Lars Backstrom, Eric Sun & Cameron Marlow.

*Find me if you can: improving geographical prediction with social and spatial proximity.*

In Proceedings of WWW 2010, Raleigh, North Carolina, USA, 2010.



Jacob G Foster, David V Foster, Peter Grassberger & Maya Paczuski.

*Edge direction and the structure of networks.*

Proceedings of the National Academy of Sciences, vol. 107, no. 24, pages 10815–10820, 2010.



A. Kaltenbrunner, G. Gonzalez, R. Ruiz de Querol & Y. Volkovich.

*Comparative analysis of articulated and behavioural social networks in a social news sharing website.*

New Review of Hypermedia and Multimedia, vol. 17, no. 3, pages 243–266, 2011.



Andreas Kaltenbrunner, Pablo Aragón, David Laniado & Yana Volkovich.

*Not All Paths Lead to Rome: Analysing the Network of Sister Cities.*

In Self-Organizing Systems, volume 8221 of *Lecture Notes in Computer Science*, pages 151–156. Springer Berlin Heidelberg, 2014.



# Bibliography II



D. Laniado, C. Castillo, A. Kaltenbrunner & M. Fuster-Morell.

*Emotions and dialogue in a peer-production community: the case of Wikipedia.*

In Proceedings of the 8th International Symposium on Wikis and Open Collaboration (WikiSym'12). ACM, 2012.



David Liben-Nowell, Jasmine Novak, Ravi Kumar, Prabhakar Raghavan & Andrew Tomkins.

*Geographic routing in social networks.*

PNAS, vol. 102, no. 33, pages 11623–11628, August 2005.



J. J. Neff, D. Laniado, K. E. Kappler, Y. Volkovich, P. Aragón & A. Kaltenbrunner.

*Jointly They Edit: Examining the Impact of Community Identification on Political Interaction in Wikipedia.*

PLoS ONE, vol. 8, no. 4, page e60584, 2013.



M.E.J. Newman.

*Mixing patterns in networks.*

Physical Review E, vol. 67, no. 2, page 26126, 2003.



John Q. Stewart.

*An Inverse Distance Variation for Certain Social Influences.*

Science, vol. 93, no. 2404, pages 89–90, 1941.



Y. Volkovich, S. Scellato, D. Laniado, C. Mascolo & A. Kaltenbrunner.

*The length of bridge ties: structural and geographic properties of online social interactions.*

In ICWSM-12 - 6th International AAAI Conference on Weblogs and Social Media. The AAAI Press, 2012.

