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

David Laniado*, †, Riccardo Tasso†, Yana Volkovich* & Andreas Kaltenbrunner*

*Barcelona Media - Innovation Center
Information, Technology & Society Research Group

†Politecnico di Milano
Dipartimento di Elettronica e Informazione

ICWSM 2011, Barcelona, July 20th
1. Introduction

2. Data extraction

3. Social interaction networks
   - Basic network parameters
   - Directed assortativity

4. Discussion trees
   - Statistics about discussions
   - Differences between macro-categories

5. Conclusions
Introduction

Wikipedia visible side

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

1 Transition period
2 Inauguration
3 First 100 days
   3.1 Expectations
   3.2 Legislation and executive orders
4 Approval ratings and opinion
   4.1 2009
   4.1.1 Handling of the economy
   4.2 2010
5 Major legislation
   5.1 Legislation signed
   5.1.1 2009
   5.1.2 2010
   5.1.3 2011
6 Personnel
   6.1 Cabinet appointees
   6.2 Notable non-Cabinet positions
   6.3 Judicial nominees
   6.3.1 Supreme Court
   6.3.2 Other courts
7 Policies
   7.1 Economy
   7.2 Ethics
     7.2.1 Lobbying reform
     7.2.2 Transparency
   7.3 Foreign policy
     7.3.1 Guantánamo Bay detention camp
     7.3.2 Overseas Contingency Operation
   7.4 Gun control
   7.5 Science and technology
   7.5.1 Cybersecurity
   7.5.2 Environment
   7.5.3 NASA
   7.5.4 Stem cell research

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)
Hollywood, Hawaii, United States[1]

Birth name Barack Hussein Obama

Nationality American

Political party Democratic

Spouse(s) Michelle Obama (m. 1992)

Children Malia Ann (b. 1990)
Natalia (Sasha) (b. 2001)

Residence The White House

Aima mater Occidental College
          Columbia University
          (B.A.)
Introduction

Article talk pages

Laniado, Tasso, Volkovich & Kaltenbrunner (Barcelona Media, Politecnico di Milano)

When the Wikipedians talk

20-07-2011 5 / 29
Introduction
The hidden side of Wikipedia

- Since 2007, growth of Wikipedia has notably slowed down [B. Suh et al.; “The singularity is not near: slowing growth of Wikipedia;” in Proc. of WikiSym ’09]

- The hidden side of Wikipedia is gaining importance
  - article talk pages → explicit coordination and discussion
  - user talk pages → personal communications (sort of public inbox)

- Article “Barack Obama”:
  - discussion split into 72 pages
  - 22,000 comments in the article talk pages
  - (17,500 edits done to the article)
Introduction

Goal
Detect patterns of interaction in the communications between the Wikipedians

- Unlike in other online discussion spaces, in Wikipedia the users discuss to reach consensus and to coordinate their activity with each other

Contribution
- extract the discussions’ structure
- analyze interaction networks
- study discussion trees
Outline

1. Introduction

2. Data extraction

3. Social interaction networks
   - Basic network parameters
   - Directed assortativity

4. Discussion trees
   - Statistics about discussions
   - Differences between macro-categories

5. Conclusions
I think the bot is archiving this talk page too soon and too frequently. [edit]

What do you think? Grundle2600 (talk) 22:53, 23 June 2009 (UTC)

I think this page is very short, so no harm in keeping the discussions a bit longer. As an incremental step I'll increased the archive time from 7 days to 10 days, per your comment. Wikidemon (talk) 01:19, 24 June 2009 (UTC)

Thank you. I didn't even know you could change it like that! Grundle2600 (talk) 02:40, 24 June 2009 (UTC)

Coup d'etat in Honduras [edit]

Why does this article not mention Obama's first coup d'etat. Just go to http://www.globalresearch.ca to find out more information about this current event. —Preceding unsigned comment added by 99.255.173.155 (talk) 03:26, 5 July 2009 (UTC)

Because there needs to be better sourcing for something as controversial as this. QueenofBattle (talk) 19:39, 5 July 2009 (UTC)

Better sourcing! Grundle2600 (talk) 01:22, 8 July 2009 (UTC)

Typing in Obama's name, caption, and the guys name does not mean better sourcing. Simply put, there has to be a very reliable source that says exactly that this was Obama's first coup d'etat. There can be no WP:OR or WP:SYNTH.

Brothejr (talk) 01:30, 8 July 2009 (UTC)

OK. That's a good point. ABC News says that Honduras' President Manuel Zelaya tried to give himself another term, despite the fact that Honduras' constitution prohibits such a thing, and that Obama said he supports Zelaya in this action. Grundle2600 (talk) 01:37, 8 July 2009 (UTC)

And of course it says much more than that - you leave out a great deal of context, not to mention leaving to the side more in-depth coverage in any number of other sources. Also you should explain exactly how you envision the Honduran coup being mentioned in this article. Personally I think it's rather important in a global history sense, but I'm not sure it deserves mention in the article on the Presidency of Barack Obama. He has talked about it, of course, but then again so has most every major leader and foreign policy NGO in the Western Hemisphere.

--Bigtimepeace | talk | contribs 07:35, 8 July 2009 (UTC)

I was just helping out with sources. I don't necessarily think that it should, or shouldn't, be cited in this article. Perhaps it could be cited in the article about the event itself, instead. Grundle2600 (talk) 08:54, 8 July 2009 (UTC)

2009 Honduran coup d'état already says, "President Barack Obama of the United States said "We believe that the coup was not legal and that President Zelaya remains the President of Honduras."[55][157]" I think that's enough, and it doesn't necessarily have to be added to this article. It could be cited in this article to, but I don't give it high priority in this article. Grundle2600 (talk) 09:04, 8 July 2009 (UTC)
Data extraction
Discussion tree for article “Presidency of Barack Obama”

- red \rightarrow root (the article)
- blue \rightarrow structural nodes
- green \rightarrow anonymous comments
- grey \rightarrow registered comments
### Dataset

Basic quantities of the data analysed


<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>#articles</strong></td>
<td>3,210,039</td>
<td></td>
</tr>
<tr>
<td><strong>#edits of article pages</strong></td>
<td>4,028,516,686</td>
<td></td>
</tr>
<tr>
<td><strong>#articles with talk page (ATP)</strong></td>
<td>871,485</td>
<td>(27.1%)</td>
</tr>
<tr>
<td><strong>#comments in ATP</strong></td>
<td>9,421,976</td>
<td></td>
</tr>
<tr>
<td><strong>#users who comment articles</strong></td>
<td>350,958</td>
<td>(2.8%)</td>
</tr>
<tr>
<td><strong>#registered users</strong></td>
<td>12,651,636</td>
<td></td>
</tr>
<tr>
<td><strong>#user talk pages (UTP)</strong></td>
<td>1,662,818</td>
<td>(13.1%)</td>
</tr>
<tr>
<td><strong>#comments in UTP</strong></td>
<td>13,670,980</td>
<td></td>
</tr>
</tbody>
</table>
1. Introduction

2. Data extraction

3. Social interaction networks
   - Basic network parameters
   - Directed assortativity

4. Discussion trees
   - Statistics about discussions
   - Differences between macro-categories

5. Conclusions
Social interaction networks

Network construction

- **article reply network** → direct replies in articles discussion pages.

- **user reply network** → direct replies in user talk pages.

- **wall network** → personal messages posted on another user’s talk page.
Basic network parameters

Global measures of the Wikipedia discussion and talk network

<table>
<thead>
<tr>
<th>variable</th>
<th>reply-NW</th>
<th>talk-NW</th>
<th>wall-NW</th>
</tr>
</thead>
<tbody>
<tr>
<td>#nodes with edges</td>
<td>204 017</td>
<td>114 258</td>
<td>1 861 702</td>
</tr>
<tr>
<td>w. in-degree ≥1</td>
<td>121 682</td>
<td>103 147</td>
<td>1 832 168</td>
</tr>
<tr>
<td>w. out-degree ≥1</td>
<td>182 881</td>
<td>63 334</td>
<td>177 331</td>
</tr>
<tr>
<td>#edges</td>
<td>1 489 734</td>
<td>852 065</td>
<td>4 412 212</td>
</tr>
<tr>
<td>mean distance</td>
<td>4.10</td>
<td>3.86</td>
<td>4.06</td>
</tr>
<tr>
<td>maximal distance</td>
<td>15</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>mean in/out-degree</td>
<td>7.30</td>
<td>7.46</td>
<td>2.37</td>
</tr>
<tr>
<td>clustering coeff.</td>
<td>0.083</td>
<td>0.053</td>
<td>0.035</td>
</tr>
<tr>
<td>reciprocity</td>
<td>0.44</td>
<td>0.45</td>
<td>0.15</td>
</tr>
</tbody>
</table>

When the Wikipedians talk
Directed assortativity

Assortativity: measure of *diversity* in networks
- do nodes having many connections preferentially interact with one another or with poorly connected nodes?

Directed assortativity: correlation between *in*- and *out*- degree of *source* and *target* nodes [J. Foster et al.; “Edge direction and the structure of networks”; in *PNAS*, 2010]

Results depend on the network structure
- To assess statistical significance, contrast the results with an ensemble of N=100 randomized networks having the same degree sequence

Normalization of the four assortativity measures for a network (Assortativity Significance Profile)
Directed assortativity profiles

Comparison of the directed Assortativity Significance Profile

Where ASP score is not significant (|Z| < 2), it is marked with the corresponding symbol at the figure bottoms.
Statistics about discussions

Distributions of users and comments by article discussion

- 85% of articles have \( \leq 10 \) comments
- 15,000 articles have \( \geq 100 \) comments
Discussion trees

Size of the discussions

- Number of users involved
- Number of *chains* of length $\geq 3$, or consecutive replies between two users
  - example chain of length 3: $A \leftarrow B \leftarrow A$
  - good indicator of *conflictive* discussions

<table>
<thead>
<tr>
<th>#articles</th>
<th>#chains</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
</tr>
</tbody>
</table>

$\text{data}$

- truncated LN: $\mu=-6.5, \sigma=3.0$
- $p$-value = 0.997

$\text{PL: } \alpha = 2.23$
- $p$-value = 0.087
- cut-off: $x_{\min} = 21$
- 95.75% discarded
Discussion trees
Length of the discussions

- **Maximum depth**
  - sensitive to the presence of individual threads

- **h-index** of the discussion tree
  - maximum number $h$ such that there are:
    $\geq h$ comments at depth $\geq h$

---

![Graphs showing the distribution of maximum depth and h-index for all discussions and discussions >100 nodes.](image-url)
## Most discussed Wikipedia articles

Top 20 articles ordered by number of chains in the discussion

<table>
<thead>
<tr>
<th>#</th>
<th>Title</th>
<th>chains</th>
<th>comments</th>
<th>users</th>
<th>h-index</th>
<th>max. depth</th>
<th>edits</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Intelligent design</td>
<td>2413</td>
<td>22454 (3)</td>
<td>954 (13)</td>
<td>16 (20)</td>
<td>20 (358)</td>
<td>9179 (53)</td>
</tr>
<tr>
<td>2</td>
<td>Gaza War</td>
<td>2358</td>
<td>17961 (6)</td>
<td>607 (47)</td>
<td>19 (2)</td>
<td>27 (28)</td>
<td>11499 (29)</td>
</tr>
<tr>
<td>3</td>
<td>Barack Obama</td>
<td>2301</td>
<td>22756 (2)</td>
<td>2360 (2)</td>
<td>18 (6)</td>
<td>21 (245)</td>
<td>17453 (6)</td>
</tr>
<tr>
<td>4</td>
<td>Sarah Palin</td>
<td>2182</td>
<td>19634 (4)</td>
<td>1221 (9)</td>
<td>17 (10)</td>
<td>25 (56)</td>
<td>12093 (24)</td>
</tr>
<tr>
<td>5</td>
<td>Global warming</td>
<td>2178</td>
<td>19138 (5)</td>
<td>1382 (5)</td>
<td>17 (10)</td>
<td>20 (358)</td>
<td>14074 (15)</td>
</tr>
<tr>
<td>6</td>
<td>Main Page</td>
<td>2065</td>
<td>32664 (1)</td>
<td>5969 (1)</td>
<td>15 (34)</td>
<td>22 (169)</td>
<td>4003 (674)</td>
</tr>
<tr>
<td>7</td>
<td>Chiropractic</td>
<td>1772</td>
<td>13684 (13)</td>
<td>243 (389)</td>
<td>18 (6)</td>
<td>29 (17)</td>
<td>6190 (204)</td>
</tr>
<tr>
<td>8</td>
<td>Race and intelligence</td>
<td>1764</td>
<td>13790 (12)</td>
<td>410 (126)</td>
<td>17 (10)</td>
<td>24 (74)</td>
<td>7615 (100)</td>
</tr>
<tr>
<td>9</td>
<td>Anarchism</td>
<td>1589</td>
<td>14385 (9)</td>
<td>496 (76)</td>
<td>20 (1)</td>
<td>28 (22)</td>
<td>12589 (19)</td>
</tr>
<tr>
<td>10</td>
<td>British Isles</td>
<td>1556</td>
<td>12044 (16)</td>
<td>576 (56)</td>
<td>17 (10)</td>
<td>23 (113)</td>
<td>4047 (658)</td>
</tr>
<tr>
<td>11</td>
<td>CRU hacking incident</td>
<td>1551</td>
<td>11536 (17)</td>
<td>474 (88)</td>
<td>17 (10)</td>
<td>20 (358)</td>
<td>2346 (2364)</td>
</tr>
<tr>
<td>12</td>
<td>Jesus</td>
<td>1397</td>
<td>17916 (7)</td>
<td>1239 (7)</td>
<td>13 (119)</td>
<td>16 (1383)</td>
<td>17081 (7)</td>
</tr>
<tr>
<td>13</td>
<td>Circumcision</td>
<td>1356</td>
<td>10469 (21)</td>
<td>436 (113)</td>
<td>17 (10)</td>
<td>26 (42)</td>
<td>7354 (117)</td>
</tr>
<tr>
<td>14</td>
<td>Homeopathy</td>
<td>1323</td>
<td>13509 (14)</td>
<td>516 (68)</td>
<td>17 (10)</td>
<td>25 (56)</td>
<td>6902 (151)</td>
</tr>
<tr>
<td>15</td>
<td>George W. Bush</td>
<td>1281</td>
<td>15257 (8)</td>
<td>1969 (3)</td>
<td>14 (65)</td>
<td>18 (676)</td>
<td>32314 (1)</td>
</tr>
<tr>
<td>16</td>
<td>September 11 attacks</td>
<td>1250</td>
<td>13830 (11)</td>
<td>1244 (6)</td>
<td>16 (20)</td>
<td>26 (42)</td>
<td>11086 (30)</td>
</tr>
<tr>
<td>17</td>
<td>Evolution</td>
<td>1165</td>
<td>13404 (15)</td>
<td>942 (16)</td>
<td>13 (119)</td>
<td>23 (113)</td>
<td>9780 (44)</td>
</tr>
<tr>
<td>18</td>
<td>Catholic Church</td>
<td>1162</td>
<td>14104 (10)</td>
<td>620 (43)</td>
<td>15 (34)</td>
<td>18 (676)</td>
<td>14082 (14)</td>
</tr>
<tr>
<td>19</td>
<td>Cold fusion</td>
<td>1098</td>
<td>8354 (29)</td>
<td>359 (174)</td>
<td>15 (34)</td>
<td>20 (358)</td>
<td>4320 (557)</td>
</tr>
<tr>
<td>20</td>
<td>2008 South Ossetia war</td>
<td>1075</td>
<td>10596 (20)</td>
<td>853 (20)</td>
<td>17 (10)</td>
<td>23 (113)</td>
<td>9930 (43)</td>
</tr>
</tbody>
</table>

In parenthesis: rank according to the corresponding variable

---

1 Climatic Research Unit
Assigning articles to macro-categories

Approach by [Kittur et al.; “What’s in Wikipedia?: mapping topics and conflict using socially annotated category structure.” In Proc. of CHI ’09]

- 21 macro-categories from Wikipedia
  - ~ direct sub-categories of Category:Main_topic_classifications
- assign each category to the closest macro-category

for each article:
- consider all categories to which it is directly assigned
- combine macro-categories corresponding to these categories to assign a percentage

improved effectiveness by adding a weight to penalise links in the “wrong direction” in the category graph.
Assigning articles to macro-categories

Example:

- Article "Barcelona" belongs to 3 categories:
  - Geography of Catalonia $\rightarrow$ Geography
  - Phoenician colonies in Spain $\rightarrow$ History
  - 10s BC establishments $\rightarrow$ History

- Barcelona $\rightarrow$ History (66.7%), Geography (33.3%)
Article discussions categorisation

% of articles with discussions in different categories

- Arts: 0.6%
- Computing: 0.5%
- Philosophy: 1.0%
- Health: 1.2%
- Belief: 2.0%
- Mathematics: 2.0%
- Language: 1.9%
- Science: 1.6%
- Business: 1.6%
- Environment: 1.2%
- Law: 1.0%
- Education: 3.1%
- Tec. & app. sci.: 3.2%
- Politics: 2.4%
- Society: 2.4%
- Sports: 3.4%
- Agriculture: 8.1%
- People: 12.1%
- Culture: 23.4%
- History & events: 23.3%
- Geogr. & places: 23.3%

When the Wikipedians talk

Laniado, Tasso, Volkovich & Kaltenbrunner (Barcelona Media, Politecnico di Milano)

20-07-2011
Article discussions by semantic area
Differences between Categories

# of chains vs. # of edits

# of users vs. h-index

Gray areas → 95% confidence interval with bootstrap test
Outline

1. Introduction
2. Data extraction
3. Social interaction networks
   - Basic network parameters
   - Directed assortativity
4. Discussion trees
   - Statistics about discussions
   - Differences between macro-categories
5. Conclusions
Conclusions

- Considerable effort for the data extraction
  - the parser will be released as free software

- Patterns in the interaction among the Wikipedians
  - strong relationship between the core and the periphery of the community
  - the existence of different roles could be further investigated

- Measures to characterize the discussions
  - h-index allows to study the depth of the discussion trees
  - reply chains help to detect conflictive topics

- Evidence of structural differences among discussions from different macro-categories


