Beyond the individual: Applications of social network analysis for the social sciences

Dr Scott A. Hale
Oxford Internet Institute
http://www.scotthale.net/, @computermacgyve

8 March 2017
The Oxford Internet Institute is . . .

- a multidisciplinary department
- of the University of Oxford
- studying “life online”
The Oxford Internet Institute is . . .

- a multidisciplinary department
- of the University of Oxford
- studying “life online”

I am . . .

- a senior data scientist / computer scientist
- analyzing large-scale data and
- conducting experiments
- to investigate multilingualism and collective action
Surveys ask questions of individuals, and we then infer large-scale / population-level characteristics from the data. This is atomistic and ignores the social connections and influences people have on one another.

Social network analysis tries to capture the social context of individuals and link micro-level behaviours to emergent micro-level outcomes.
Classic findings

- Weak ties
- Homophily
- Six degrees of separation
Networks (graphs) are set of nodes (vertices) connected by edges (links, ties, arcs).
Networks (graphs) are set of nodes (vertices) connected by edges (links, ties, arcs).

Further details:
- Whole vs. ego: whole networks have all nodes within a natural boundary (platform, organization, etc.). An ego network has one node and all of its immediate neighbors.
- Edges can be directed or undirected and weighted or unweighted.
- Additionally, networks may be multilayer and/or multimodal.
Why?

- Characterize network structure
  - How far apart / well-connected are nodes?
  - Are some nodes at more important positions?
  - Is the network composed of communities?
- How does network structure affect processes?
  - Information diffusion
  - Coordination/cooperation
  - Resilience to failure/attack
A network

First questions when approaching a network

- What are edges? What are nodes?
- What kind of network?
- Inclusion/exclusion criteria
Network data formats

Data format
The simplest form of network is simply two spreadsheets or comma-separated values (csv) files. One sheet for the nodes of the network and one sheet for the edges of the network.

Example

Nodes
id  name  profession
n1  Alan  student
n2  Betty  teacher
n3  Charlotte  student

Edges
id  source  target  weight
e1  n1  n2  1
e2  n1  n3  100
Gephi

- Open-source, cross-platform GUI interface
- Primary strength is to visualize networks
- Basic statistical properties are also available
- Alternatives include NodeXL, Pajek, GUESS, NetDraw, Tulip, and more
Other software

- **NodeXL** ([http://nodexl.codeplex.com/](http://nodexl.codeplex.com/)):
  - Add-in for Microsoft Excel (Windows only)
  - Good data collection options (Twitter, Facebook, YouTube, ...)
  - Basic visualization

- **Pajek**
  - [http://pajek.imfm.si/doku.php?id=download](http://pajek.imfm.si/doku.php?id=download)
  - Standalone, Windows (or Linux with wine)
  - Good interactive environment for metrics and basic visualization

- **iGraph**
  - `install.packages("igraph")` in R ([r-project.org](http://r-project.org)) (or python)
  - Cross-platform
  - Text driven: powerful for analysis

- **NetworkX**
  - Cross-platform python library
  - Examples/introduction in this slide deck
An example: bilingualism

Given the big differences in the information available in different languages, do bilingual users on social media serve as ‘bridges’ introducing content from one language into another language online?
Bilingualism common (offline)

“multilingualism...[is] the norm for most of the world’s societies” (Birner, 2005), with over half of Europe and over a fifth of the US multilingual (Erard, 2012); yet, many platforms are designed only with monolingual users in mind.

Novel information

Benefits for connecting clusters both at a large, network level & at an individual level (Granovetter, 1973; Burt, 2004; Uzzi & Spiro, 2005; Aral & Alstyne, 2011). Beyond ‘filter bubbles’ (Pariser, 2011), language is a large restriction on available content.
Twitter: Language and structure

- Twitter user

○ Twitter user
Twitter: Language and structure

- Twitter user
  - Mention, Retweet

Clusters: Automated groups of strongly-connected users based purely on network structure

Colour: Most-used language
Twitter: Language and structure

- Twitter user
- Mention, Retweet
- Clusters: Automated groups of strongly-connected users based purely on network structure
Twitter: Language and structure

- Twitter user
- Mention, Retweet
- Clusters: Automated groups of strongly-connected users based purely on network structure

Colour: Most-used language
Label propagation algorithm (Raghavan, Albert, & Kumara, 2007) found 20,253 clusters.

Histograms of the size of clusters (left) and the number of languages within each cluster (right). Modularity score of 0.81 for this division of the network.
Twitter: Bilinguals bridge clusters

- Twitter user
- Mention, Retweet
- Colour: Most-used language
Twitter: Bilinguals bridge clusters

- Twitter user
- Mention, Retweet

Colour: Monolingual language; Black: Bilingual
Twitter: Bilinguals bridge clusters

- Twitter user

→ Mention, Retweet

Colour: Monolingual language;
Black: Bilingual
Size of the largest, weakly-connected component (left), total number of components (center), and average size of the components (right) created by removing all bilingual users, an equivalent number of monolingual users randomly, an equivalent number of all users randomly, and removing all bilingual users from a network with the same degree distribution but with edges randomly shuffled. Box plots show values from 100 realizations. Mean values are indicated with +.
Removing bilingual users results in a smaller largest, weakly-connected component compared to removing an equivalent number of monolingual users randomly, an equivalent number of all users randomly, or removing all bilingual users from a network with the same degree distribution but with edges randomly shuffled. Box plots show values from 100 realizations. Mean values are indicated with +.
The percentage of remaining users not in the largest connected component after removing all users in different languages.
Twitter: Cross-language connections

Mentions and retweets across languages

- Nodes represent most-used language (size proportional to number of primary users)
- Edges are weighted by the percent error in the expected vs. the actual number of mentions and retweets between languages
- Colors indicate clusters found by the infomap community detection algorithm
Bilingual bridging

Bilinguals on Twitter are in network positions to serve as bridges,

but what about the content they write in different languages?

Bilinguals able to recall memories and taught knowledge more easily when the language of encoding matched the language of retrieval (Marian & Neisser, 2000; Marian & Fausey, 2006)
Similar text in different languages
Barack Obama

From Wikipedia, the free encyclopedia

(Redirected from Obama)

"Barack" and "Obama" redirect here. For his father, see Barack Obama Sr. For other uses of "Barack", see Barack (disambiguation). For other uses of "Obama", see Obama (disambiguation).

Barack Hussein Obama II (US-1/baˈræk hooˈsen əbəˈmaː/;[1][2] born August 4, 1961) is an American politician who is the 44th President of the United States. He is the first African American to hold the office and the first president born outside the continental United States. Born in Honolulu, Hawaii, Obama is a graduate of Columbia University and Harvard Law School, where he was president of the Harvard Law Review. He was a community organizer in Chicago before earning his law degree. He worked as a civil rights attorney and taught constitutional law at the University of Chicago Law School between 1992 and 2004. While serving three terms representing the 13th District in the Illinois Senate from 1997 to 2004, he ran unsuccessfully in the Democratic primary for the United States House of Representatives in 2000 against incumbent Bobby Rush.

In 2004, Obama received national attention during his campaign to represent Illinois in the United States Senate with his victory in the March Democratic Party primary, his keynote address at the Democratic National Convention in July, and his election to the Senate in November. He began his presidential campaign in 2007 and, after a close primary campaign against Hillary Clinton in 2008, won sufficient delegates in the Democratic Primaries to receive the presidential nomination. He then defeated Republican nominee John McCain in the general election, and was inaugurated as president on January 20, 2009. Nine months after his inauguration, Obama was controversially named the 2009 Nobel Peace Prize laureate.

During his first two years in office, Obama signed into law economic stimulus legislation in response to the Great Recession in the form of the American Recovery and Reinvestment Act of 2009 and the Tax Relief, Unemployment Insurance Reauthorization, and Job Creation Act of 2010. Other major domestic initiatives in his first term included the Patient Protection and Affordable Care Act, often referred to as "Obamacare"; the Dodd–Frank Wall Street Reform and Consumer Protection Act; and the Don't Ask, Don't Tell Repeal Act of 2010. In foreign policy, Obama ended U.S. military involvement in the Iraq War, increased U.S. troop levels in Afghanistan, signed the START arms control treaty with Russia, ordered U.S. military involvement in Libya in opposition to Muammar Gaddafi, and ordered the military operation that resulted in the death of Osama bin Laden. In January 2011, the Republicans regained control of the House of Representatives as the Democratic Party lost a total of 63 seats; and, after a lengthy debate over federal spending and whether or not to raise the nation's debt limit, Obama signed the Budget Control Act of 2011 and the American Taxpayer Relief Act of 2012.

Obama was reelected president in November 2012, defeating Republican nominee Mitt Romney, and was sworn in for a second term on January 20, 2013. During his second term, Obama has promoted domestic policies related to gun control in response to the Sandy Hook Elementary School shooting, and has called for greater inclusiveness for LGBT Americans, while his administration has filed briefs which urged the Supreme Court to strike down part of the federal Defense of Marriage Act (United States v. Windsor) and state level same-sex marriage bans (Obergefell v. Hodges) as unconstitutional. In foreign policy, Obama ordered U.S. military intervention in Iraq in response to gains made by ISIL after the 2011 withdrawal from Iraq, continued the process of ending U.S. combat operations in Afghanistan, promoted discussions that
Barack Obama Sr. (born August 4, 1961) is an American politician who is the first African American to hold the office and the first president born in the Hawaiian Islands. Born in Honolulu, Hawaii, Obama is a graduate of Columbia University and Harvard Law School, where he was a community organizer in Chicago before earning a law degree and teaching constitutional law at the University of Chicago Law School for three terms representing the 13th District in the Illinois Senate from 1997 to 2004. At age 47, he was elected to the United States House of Representatives in 2000 against incumbent Tony G. Coelho.

During his campaign to represent Illinois in the United States Senate with his keynote address at the Democratic National Convention in July, and his victory speech at Grant Park, he began his presidential campaign in 2007 and, after a close primary campaign, won sufficient delegates in the Democratic Party primaries to receive the presidential nomination of John McCain in the general election, and was inaugurated as president of the United States on January 20, 2009. This was the first time a black person had been elected to the presidency of the United States.

Obama signed into law economic stimulus legislation in response to the Great Recession and Reinvestment Act of 2009 and the Tax Relief, Unemployment Insurance Reauthorization, and Job Creation Act of 2010. Other major domestic initiatives in his first term included the Patient Protection and Affordable Care Act, which was referred to as “Obamacare”; the Dodd–Frank Wall Street Reform and Consumer Protection Act of 2010. In foreign policy, Obama ended U.S. military involvement in Afghanistan, signed the New START arms control treaty with Russia, ordered opposition to Muammar Gaddafi, and ordered the military operation that resulted in his approval rating for the Budget Control Act of 2011 and the American Taxpayer Relief Act of 2012.

In foreign policy, Obama ordered U.S. military intervention in Iraq in response to gains made by ISIL after the 2011 withdrawal from Iraq, continued the process of ending U.S. combat operations in Afghanistan, promoted discussions that
A large number of users did not edit any of the same articles in their primary languages, but a large number of users also always edited the same articles in their primary languages.
A large number of users did not edit any of the same articles in their primary languages, but a large number of users also always edited the same articles in their primary languages.
Awareness and discovery of relevant foreign-language content is difficult for bilinguals.

Many interfaces are designed for monolingual speakers.
Other network statistics

With many nodes visualizations are often difficult/impossible to interpret. Statistical measures can be very revealing, however.

- **Node-level**
  - Degree (in, out): How many incoming/outgoing edges does a node have?
  - Centrality (next slide)
  - Constraint

- **Network-level**
  - Components: Number of disconnected subsets of nodes
  - Density: \( \frac{\text{observed edges}}{\text{maximum number of edges possible}} \)
  - Clustering coefficient \( \frac{\text{closed triplets}}{\text{connected triples}} \)
  - Path length distribution
  - Distributions of node-level measures
Centrality measures

- **Degree**
- **Closeness**: Measures the average geodesic distance to ALL other nodes. Informally, an indication of the ability of a node to diffuse a property efficiently.
- **Betweenness**: Number of shortest paths the node lies on. Informally, the betweenness is high if a node bridges clusters.
- **Eigenvector**: A weighted degree centrality (inbound links from highly central nodes count more).
- **PageRank**: Not strictly a centrality measure, but similar to eigenvector but modeled as a random walk with a teleportation parameter
Implications: Bilingualism

Bilingualism is the norm offline, but not online
- Twitter: 11%, Wikipedia: 15%
- Differing level of competency in different languages (mirrors offline)
- Proportion of bilinguals $\propto \frac{1}{\text{self-focus bias}}$
- Importance of content discovery and interface design

Bilinguals are important
- Small, dedicated group; large-sized contributions in primary languages
- Contributions in non-primary language small, but valuable
- Different types of contributions
- Mixed evidence on serving as bridges between languages
Active areas of research

- Multilingual topic modelling, word embeddings
- (Word-level) language detection
- Identifying similar/translated content across languages

Computational sociolinguistics: Nguyen, Doğruöz, Rosé, and de Jong (2016)
Applying results

Helicopter crash at Okinawa International University

Overview

At approximately 2:15pm on August 13, 2004, a CH-53D heavy-lift transport helicopter
Applying results

Bridge
Translate the Global Web

Bridge: Social Translation for Social Media. Follow us on Facebook and Twitter.

Translated by John Doe
Emirates minister of state Sultan Aljaber: the Egyptian people assured their ability to defy challenges.

Translated by Nora Younis
Fish, milk, tamarind. #FishAndMilkAndTamarind is a literal translation of a common Egyptian expression that emphasizes ingenuity. In the case between the various "variations" of the president. Often translated as "knowledge" or "wisdom".

Translated by Tom Trenchard
A civilian, elected president wearing military uniform on a royal yacht. Fish, milk, tamarind.

Make sense of the global web.
Further resources for networks

- NetworkX documentation (http://networkx.github.io/documentation/latest/reference/)
- Newman, M.E.J., Networks: An Introduction
- Kadushin, C., Understanding Social Networks: Theories, Concepts, and Findings
- De Nooy, W., et al., Exploratory Social Network Analysis with Pajek
- Shneiderman B., and Smith, M., Analyzing Social Media Networks with NodeXL
Beyond the individual:
Applications of social network analysis for the social sciences

Dr Scott A. Hale
Oxford Internet Institute
http://www.scotthale.net/, @computermacgyve

8 March 2017


