UNDERSTANDING OFFLINE POLITICAL SYSTEMS BY MINING ONLINE POLITICAL DATA

David Lazer
Northeastern & Harvard
d.lazer@neu.edu

Oren Tsur
Northeastern & Harvard
orents@seas.harvard.edu

Tina Eliassi-Rad
Northeastern & Rutgers
tina@eliassi.org
# Outline

<table>
<thead>
<tr>
<th>Time</th>
<th>Presenter</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00 - 9:50</td>
<td>David</td>
<td>Political inquiry, new science of politics, exemplary data</td>
</tr>
<tr>
<td>9:50 - 10:30</td>
<td>Oren</td>
<td>Exponential Random Graph Models</td>
</tr>
<tr>
<td>10:30 - 11:00</td>
<td>Coffee Break</td>
<td>Coffee Break</td>
</tr>
<tr>
<td>11:00 - 11:20</td>
<td>Oren</td>
<td>Networks of political figures on Twitter</td>
</tr>
<tr>
<td>11:20 - 11:50</td>
<td>Tina</td>
<td>Roles in socio-political networks</td>
</tr>
<tr>
<td>11:50 - 12:00</td>
<td>David</td>
<td>Wrap-up &amp; questions</td>
</tr>
</tbody>
</table>
## Outline

<table>
<thead>
<tr>
<th>Time</th>
<th>Speaker</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00 - 9:50</td>
<td>David</td>
<td>Political inquiry, new science of politics, exemplary data</td>
</tr>
<tr>
<td>9:50 - 10:30</td>
<td>Oren</td>
<td>Exponential Random Graph Models</td>
</tr>
<tr>
<td>10:30 - 11:00</td>
<td></td>
<td>Coffee Break</td>
</tr>
<tr>
<td>11:00 - 11:20</td>
<td>Oren</td>
<td>Networks of political figures on Twitter</td>
</tr>
<tr>
<td>11:20 - 11:50</td>
<td>Tina</td>
<td>Roles in socio-political networks</td>
</tr>
<tr>
<td>11:50 - 12:00</td>
<td>David</td>
<td>Wrap-up &amp; questions</td>
</tr>
</tbody>
</table>
USING BIG DATA TO UNDERSTAND POLITICS

David Lazer
Northeastern/Harvard
Outline

• Some big themes in politics
• The opportunity to create a new science of politics
• Exemplar data
• Cautionary tales
Big themes in politics

• Collective action
• Political communication
• Power
The paradox of collective action
The paradox of collective action

- Social movements
- Voting
- Contributing to campaigns
- Vaccination
Why?

- “Selective benefits/penalties”
- Solidarity
- Norms
Political communication

THE GETTYSBURG ADDRESS

November 19, 1863
At the Dedication of the Soldiers’ National Cemetery in Gettysburg, Pennsylvania:

Four score and seven years ago our fathers brought forth on this continent, a new nation, conceived in Liberty, and dedicated to the proposition that all men are created equal.

Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. We are met on a great battle-field of that war. We have come to dedicate a portion of that field, as a final resting place for those who here gave their lives that that nation might live. It is altogether fitting and proper that we should do this.

But, in a larger sense, we can not dedicate — we can not consecrate — we can not hallow — this ground. The brave men, living and dead, who struggled here, have consecrated it, far above our poor power to add or detract. The world will little note, nor long remember what we say here, but it can never forget what they did here. It is for us the living, rather, to be dedicated here to the unfinished work which they who fought here have thus far so nobly advanced. It is rather for us to be here dedicated to the great task remaining before us — that from these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion — that we here highly resolve that these dead shall not have died in vain — that this nation, under God, shall have a new birth of freedom — and that government of the people, by the people, for the people, shall not perish from the earth.

Abraham Lincoln
The construction of language in politics

• Testing of different linguistic constructions ("estate taxes" vs "death taxes")
  • Surveys, focus groups, etc
• Process of dissemination to elites
• Re-dissemination via media
Three dimensions of power (Lukes)

• **Decision making**: When A gets B to do something B otherwise would not do.
• **Agenda setting**: what is and is not discussed.
• **Normative influence**: what you think is in your interest
Time for a new science of politics

• Most social science is:
  • Static
  • Spatially and socially decontextualized
  • And small scale (hundreds or thousands of individuals)
Time for a new science of politics

• The new science of politics
  • dynamic
  • Spatially and socially embedded
  • And societal or even globally spanning
Exemplar big data

• $ in politics
• Political language
Federal elections commission data

<table>
<thead>
<tr>
<th>A.</th>
<th>Full Name (Last, First, Middle Initial)</th>
<th>Date of Receipt</th>
<th>Transaction ID</th>
<th>Amount of Each Receipt this Period</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Robert Perez</td>
<td>03/10/2008</td>
<td>C2017416</td>
<td>500.00</td>
</tr>
<tr>
<td></td>
<td>Mailing Address: One Financial Ctr., 40th Floor</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Boston, MA 02111</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FEC ID number of contributing federal political committee</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Name of Employer: Mintz Levin Colin Ferris</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Occupation: Chairman</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Receipt For: 2008</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>X Primary</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>General</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Other (specify)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Election Cycle-to-Date ▼</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>500.00</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>B.</th>
<th>Full Name (Last, First, Middle Initial)</th>
<th>Date of Receipt</th>
<th>Transaction ID</th>
<th>Amount of Each Receipt this Period</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Larry Gray</td>
<td>03/10/2008</td>
<td>C2017417</td>
<td>1000.00</td>
</tr>
<tr>
<td></td>
<td>Mailing Address: 20 Bridle Path</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Westwood, MA 02090</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FEC ID number of contributing federal political committee</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Name of Employer: Rasky Barden, Inc.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Occupation: Partner</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Receipt For: 2008</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>X Primary</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>General</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Other (specify)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Election Cycle-to-Date ▼</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1000.00</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>C.</th>
<th>Full Name (Last, First, Middle Initial)</th>
<th>Date of Receipt</th>
<th>Transaction ID</th>
<th>Amount of Each Receipt this Period</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>John Repa</td>
<td>03/10/2008</td>
<td>C2017410</td>
<td>500.00</td>
</tr>
<tr>
<td></td>
<td>Mailing Address: 58 Farnham St.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Belmont, MA 02448-2172</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FEC ID number of contributing federal political committee</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Name of Employer:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Occupation:</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
FEC data

- Contributor name
- Occupation
- Employer
- Address/zip of contributor
- Receiving committee (unique id)
- Donation amount
- Date

http://www.fec.gov/finance/disclosure/ftpdet.shtml
And inferred network among contributors

From Ruths and Lazer (2009)
Inferred relationships…

- Boston
  - Ryan Vincent & Carla Meyer - board members
- Washington DC
  - Ed Rogers & Lanny Griffith - partners in lobbying firm
- Los Angeles
  - Spielberg & Katzenberg - film producers
- NYC
  - Debra Black & Judith Hannan - board members
There are some interesting exceptions where a clique contains more than one organizational. Here we find the highest weighted cliques for two companies, Pfizer and Pharmacia Upjohn. It is not surprising that they are connected because the two companies merged nine years earlier, and the employee simply used the old name instead of the new one they probably should have. The yellow dots indicates Pfizer and the orange Pharmacia Upjohn.

We can see some sub-structure using the Singular Value Decomposition layout, where Mary Ambrow is separated from the main cluster and seems to have a different role in the company. Although not shown, she is connected to external cliques, which is why the SVD layout pulls her away from the main clique.
FYI, new $ data..

• Federal Election Commission data do not have unique identifiers…

• And disambiguation is a big barrier to doing anything with the data

• So we synthesized unique identifiers

Political language

- Myriad of sources…
- Public statements data from Votesmart
With Yu-ru Lin
osama bin laden

on the web press releases
Example

• Data
  • 0.5 million documents from public statements of Members of US Congress from Votesmart

• Computational methods
  • Tracking semantic convergence
  • Randomized n-gram extraction

Lin et al 2015
In a large corpus, multiple types of convergence

Together producing a “bumpy” distribution
in theory

in data

8-gram (jaccard) similarity between document pairs
4-gram
32-gram
Social media

• Twitter
• Facebook (?)
• Tumblr
• Anything you can scrape from the Web.
• Etc etc etc
But handle (big) data with care, a few quick lessons from the failure of Google Flu Trends

- Nobody can tell you’re a dog on the Internet— and that’s not a good thing if you are trying to understand *humans*.
  - Value of curated data— sometimes < 1% of the data is way better than 100% of the data
- Algorithmic changes— e.g., algorithmic sorting in Facebook and Twitter
- Evolving norms— example of hashtags in Twitter

*Lazer et al 2015*
Example…

• Can we classify people as liberal or conservative based on the language they use?

• **Answer:** yes, lit suggests 90+% accuracy is possible based on snapshots of language use.

• **But:** these findings turn out to be ephemeral *(Cohen and Ruths ICWSM 2013)*
## Outline

<table>
<thead>
<tr>
<th>Time</th>
<th>Speaker</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00 - 9:50</td>
<td>David</td>
<td>Political inquiry, new science of politics, exemplary data</td>
</tr>
<tr>
<td>9:50 - 10:30</td>
<td>Oren</td>
<td>Exponential Random Graph Models</td>
</tr>
<tr>
<td>10:30 - 11:00</td>
<td></td>
<td>Coffee Break</td>
</tr>
<tr>
<td>11:00 - 11:20</td>
<td>Oren</td>
<td>Networks of political figures on Twitter</td>
</tr>
<tr>
<td>11:20 - 11:50</td>
<td>Tina</td>
<td>Roles in socio-political networks</td>
</tr>
<tr>
<td>11:50 - 12:00</td>
<td>David</td>
<td>Wrap-up &amp; questions</td>
</tr>
</tbody>
</table>
## Outline

<table>
<thead>
<tr>
<th>Time</th>
<th>Speaker</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00 - 9:50</td>
<td>David</td>
<td>Political inquiry, new science of politics, exemplary data</td>
</tr>
<tr>
<td>9:50 - 10:30</td>
<td>Oren</td>
<td>Exponential Random Graph Models</td>
</tr>
<tr>
<td>10:30 - 11:00</td>
<td></td>
<td>Coffee Break</td>
</tr>
<tr>
<td>11:00 - 11:20</td>
<td>Oren</td>
<td>Networks of political figures on Twitter</td>
</tr>
<tr>
<td>11:20 - 11:50</td>
<td>Tina</td>
<td>Roles in socio-political networks</td>
</tr>
<tr>
<td>11:50 - 12:00</td>
<td>David</td>
<td>Wrap-up &amp; questions</td>
</tr>
</tbody>
</table>
EXPONENTIAL RANDOM GRAPH MODELS

Oren Tsur
Northeastern/Harvard
Available datasets
Types of “political” datasets

- **Publicly Available**
- **Proprietary**
- **Elite Users**
- **General Public**

**Publicly Available**

**Proprietary**

**Elite Users**

**General Public**
Example: “Money Talks”

- Finding network/cluster/community similarities btw. contributions networks and speech.
Networks – Informal Introduction
A political setting
Is this a random network?
Network analysis

- Given a network
  - Link prediction
  - Community detection
  - Role discovery

- Network dynamics
  - Evolution
  - Contagion, diffusion, cascades

- Network formation
  - Social factors for the above
Networks Basics
Is this a random network?
Likelihood of an observed graph

Given $|N|=12$ and $|E|=13$ (directed)

- There are $2 \cdot \binom{12}{2} = 132$ options for edge placement
- Edges are distributed independently
- So the number of possible graphs is $\binom{132}{13} = 3.22 \times 10^{17}$
- All of these graphs are equally likely...
Erdos-Renyi networks: 
A generative-probabilistic approach

• We assume **edge independence**.

• Edges are generated by a **Bernoulli process** with a parameter $p$.

• We generate a graph $G(N, p)$ by:
  
  • For each ordered pair $(u, v)$ of nodes from $N$:
    
    • $E += (u, v)$ with a probability $p$

• Each graph with $n$ nodes and $m$ edges has the following likelihood:

$$p^m \cdot (1-p)^{\binom{n}{2} - m}$$
Terminology: “Graphs” vs. “Networks”

[informal:]

- Graphs are mathematical (topological) concepts defined by nodes and edges.
- [Social] Networks represent the outcome of some social **process** (can be dynamic)
- In networks we care about **dependency** between nodes and edges.
Networks are not “random” graphs

- **Goal**: Find a plausible (and interesting) model explaining the creation of an observed network.

- **Assumptions (for simplicity)**:
  - Observed network is *fixed*.
  - Edge formation is *not* random.
  - Network was generated based on latent factors.
  - We can speculate about the factors:
    - Common sense
    - Social science theory
    - Guess in the wild
Exponential Random Graph Models (ERGM)

- General form:
  \[ Pr(Y = y) = \left( \frac{1}{k} \right) \exp \sum_A \theta_A g_A(y) \]
- Where:
  - \( A \) is a specific “configuration” (e.g. reciprocity)
  - \( \theta_A \) is a parameter corresponding to configuration \( A \).
  - \( g_A(y) = \prod_{y_{ij} \in A} y_{ij} \) the network statistic corresponding to \( A \).
  - \( y_{ij} \in \{0,1\} \) 1 iff the \( ij \) edge is observed in \( y \).
  
  - For simplicity we generalize \( g \) and \( A \) (sum on edges inst. mult)
  - \( k \) is a normalization factor, making the general form a proper probability distribution.
Example of some model features

- **General form:** 
  \[ Pr(Y = y) = \left( \frac{1}{k} \right) \exp \sum_A \theta_A g_A(y) \]

- **Baseline model (Erdos Renyi):** 
  \[ Pr(Y = y) = \left( \frac{1}{k} \right) \exp \sum_{ij} \theta_{ij} \]

- **Examples for other terms:**
  - **Formal leadership (nodal):** 
    \[ \sum_{ij, j \in \text{Leaders}} \theta_{\text{leadership}} y_{ij} \]
  - **Reciprocity (dyad):** 
    \[ \sum_{ij} \theta_{\text{reciprocity}} y_{ij} y_{ji} \]
  - **Cyclic triad (dyad):** 
    \[ \sum_{ijk} \theta_{\text{cTriad}} y_{ij} y_{jk} y_{ki} \]

- **So “simple” toy model to estimate:**

\[
Pr(Y = y) = \left( \frac{1}{k} \right) \exp \sum_A \theta_A g_A(y) = \left( \frac{1}{k} \right) \exp \left( \sum_{ij} \theta_{ij} + \sum_{ij, j \in \text{Leaders}} \theta_{\text{leadership}} y_{ij} + \sum_{ij} \theta_{\text{reciprocity}} y_{ij} y_{ji} + \sum_{ijk} \theta_{\text{cTriad}} y_{ij} y_{jk} y_{ki} \right)
\]
Parameter estimation

- Markov Chain Monte Carlo
  - Metropolis Hastings
  - [There are other algs + new developments]

- Issues:
  - Degeneracy
  - Stability (over subsampling, incomplete networks, thresholding)
  - No direct temporal modeling
  - Not suitable for large network (estimation is problematic)
  - Interpretation needed
ERGM resources

- MCMC estimation of ERGMs
  - http://www.cmu.edu/joss/content/articles/volume3/Snijders.pdf
- R packages: statnet, network, ergm
- ERGM introduction, package documentation and examples
  - https://cran.r-project.org/web/packages/ergm/vignettes/ergm.pdf
- [New] Generalized-ERGM (+beta implementation)
- Many other tutorials, variations and examples (online)
# Outline

<table>
<thead>
<tr>
<th>Time</th>
<th>Name</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00 - 9:50</td>
<td>David</td>
<td>Political inquiry, new science of politics, exemplary data</td>
</tr>
<tr>
<td>9:50 - 10:30</td>
<td>Oren</td>
<td>Exponential Random Graph Models</td>
</tr>
<tr>
<td>10:30 - 11:00</td>
<td></td>
<td>Coffee Break</td>
</tr>
<tr>
<td>11:00 - 11:20</td>
<td>Oren</td>
<td>Networks of political figures on Twitter</td>
</tr>
<tr>
<td>11:20 - 11:50</td>
<td>Tina</td>
<td>Roles in socio-political networks</td>
</tr>
<tr>
<td>11:50 - 12:00</td>
<td>David</td>
<td>Wrap-up &amp; questions</td>
</tr>
<tr>
<td>Time</td>
<td>Speaker</td>
<td>Title</td>
</tr>
<tr>
<td>--------------</td>
<td>---------</td>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>9:00 - 9:50</td>
<td>David</td>
<td>Political inquiry, new science of politics, exemplary data</td>
</tr>
<tr>
<td>9:50 - 10:30</td>
<td>Oren</td>
<td>Exponential Random Graph Models</td>
</tr>
<tr>
<td>10:30 - 11:00</td>
<td></td>
<td>Coffee Break</td>
</tr>
<tr>
<td>11:00 - 11:20</td>
<td>Oren</td>
<td>Networks of political figures on Twitter</td>
</tr>
<tr>
<td>11:20 - 11:50</td>
<td>Tina</td>
<td>Roles in socio-political networks</td>
</tr>
<tr>
<td>11:50 - 12:00</td>
<td>David</td>
<td>Wrap-up &amp; questions</td>
</tr>
</tbody>
</table>
Data

Political Twitter
- 6 month full stream
- 158817 tweets
- Graph is directed
- Edge threshold (@>3)
- 439 Members of 114th Congress (Current)
- |E| = 9167 (after thresholding)

Joint Statements (JS)
- Full term (112th congress)
- 8979 statements
- Graph is undirected
- Edge threshold: normalized-weighted
- 435 members of 112th Congress (2011-13)
- |E| = 3188 (after thresholding)
Model features (factors/terms)

**Nodal Factors**
- Party
- Age
- Gender
- Seniority (terms in congress)
- State, region, district
- Formal leadership position
- Committee membership

**Dyad Factors**
- Reciprocity
- Cyclic triads
- Transitive triads
- Shared committee membership
- In/out-star
Members of U.S. Congress (Twitter)

6 month, 159K tweets
directed, by @, threshold=3
Members of U.S. Congress (Twitter)

6 month, 159K tweets
directed, by @, threshold=3
Members of U.S. Congress (Twitter)

One big mess (hairball)

No obvious partisanship

6 month, 159K tweets
directed, by @, threshold=3
Research questions

• What latent factors dominate link formation?
• Does network analysis fit with what we know (Political Science theory, other quant. works)?
ERGM results

- **Significant:**
  - Reciprocity matters
  - Seniority matters
  - Cyclic triads

- **Not found significant (surprising):**
  - Partisan homophily
  - Formal leadership role
  - We checked for other terms, e.g.:
    - Gender
    - State
    - Region
    - Committee membership network
ERGM results - Twitter

Significant factors in **nodal** (independent) model:
- Number of edges (Bernoulli)
- Seniority (senior members attract incoming nodes)
- **Surprising:** state, party, shared committees and formal leadership were not found significant.

Significant **dyadic** (dependent) factors:
- Reciprocity (could this be a bias of the @ mechanism?)
- 2-in-star
- Cyclic-triads
- Transitive-triads term yielded degenerate models
- Seniority significance disappeared after introducing dyadic factors
  - Probably covered by the 2-in-star
Interpretation (1)

- Seniority matters
- Reciprocity rules (in politics; in conversing; in life?)
- Leadership is not a factor (masked by seniority?!)

But:

- This is not aligned with the JS network
  - Lack of seniority is a significant factor (new members are connected)
  - Leadership is a significant factor

Why? (Is there a political scientist in the room?)

- Technical: different networks (directed, vs. undirected)
- Social 1: different networks (“wild” vs. collaborative by definition)
- Social 2: different social processes shape different network dynamics
Interpretation (cont.)

But

- Leadership (speakers, whips, majority/minority leader) has high/top centrality in relevant centrality measures (in/out/deg, betweenness)

And

- In a frame of mind (Tsur et al. ACL 2015), we find:
  - strong partisanship even in subtle topics (=framing campaigns)
  - Strong party discipline (stronger for Republicans)
Living happily ever after?!
Partisan divergence and discipline
Partisan divergence and discipline

Top 20 Hashtags per Party

- gopshutdown
- restorethevra
- takeitdown
- miicwhatsnext
- womensucceed
- standwithpp
- actonclimate
- vra50
- bringbackourgirls
- exim4jobs
- lgbt
- socialsecurity
- lwcf
- tpp
- exim
- scotus
- medicare
- popeindc
- tbt
- jobs
- cuba
- veterans
- tpa
- irandeal
- cures2015
- iran
- plannedparenthood
- obamacare
- va
- epa
- tcot
- il10
- defundpp
- tpa4usjobs
- ia03
- nonnucleariran
- prolife

% (normalized) use by party
Partisan divergence and discipline

Top 20 Hashtags per Party

- gapshutdown
- restorethervra
- takeitdown
- miicwhatsnext
- womensucceed
- standwithpp
- actonclimate
- vra50
- bringbackourgirls
- exim4jobs
- lgbt
- socialsecurity
- lwcf
- tpp
- exim
- scotus
- medicare
- popeindc
- tbt
- jobs
- cuba
- veterans
- tpa
- irandeal
- cures2015
- iran
- plannedparenthood
- obamacare
- va
- epa
- tcot
- il10
- defundpp
- tpa4usjobs
- ia03
- noncleariran
- prolif

% (normalized) use by party
Partisan divergence and discipline

Top 20 Hashtags per Party

- gapshutdown
- restorethevra
- takeitdown
- miicwhatsnext
- womensucceed
- standwithpp
- actonclimate
- vra50
- bringbackourgirls
- exim4jobs
- lgbt
- socialsecurity
- lwcf
- tpp
- exim
- scotus
- medicare
- popeindc
- tbt
- jobs
- cuba
- veterans
- tpa
- irandeal
- cures2015
- iran
- plannedparenthood
- obamacare
- va
- epa
- tcot
- il10
- defundpp
- tpa4usjobs
- ia03
- nonnucleariran
- prolif

% (normalized) use by party
Leadership and sub-communities

- Leadership is central in the Twitter network.
- Even in marginal “campaigns”
  - Involving ~20% of the party members

#smallBiz
|N|=51, |E|=74
Total: 238

#immigrationReform
|N|=43, |E|=107
Total: 78
Collaborative partisan hashtagging

Democrat use more hashtags, less effectively

Republicans have higher average uses per member in ALL ranges

In line with findings at Tsur et al. ACL 2015
Workshop Announcements

2 WS on NLP and Computational Social Science (NLP+CCS):

- **WebSci** – Hannover, Germany, May 2016  
  (deadline: March 25)
- **EMNLP** – Austin, Texas, November 2016  
  (deadline: TBA)

Politics and networks

- **Political Networks** (PolNets) – St. Louis, Missouri, June, 2016  
  Abstract based. (deadline: April 15)
<table>
<thead>
<tr>
<th>Time</th>
<th>Speaker</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00 - 9:50</td>
<td>David</td>
<td>Political inquiry, new science of politics, exemplary data</td>
</tr>
<tr>
<td>9:50 - 10:30</td>
<td>Oren</td>
<td>Exponential Random Graph Models</td>
</tr>
<tr>
<td>10:30 - 11:00</td>
<td></td>
<td>Coffee Break</td>
</tr>
<tr>
<td>11:00 - 11:20</td>
<td>Oren</td>
<td>Networks of political figures on Twitter</td>
</tr>
<tr>
<td>11:20 - 11:50</td>
<td>Tina</td>
<td>Roles in socio-political networks</td>
</tr>
<tr>
<td>11:50 - 12:00</td>
<td>David</td>
<td>Wrap-up &amp; questions</td>
</tr>
</tbody>
</table>
ROLES IN SOCIO-POLITICAL DATA

Tina Eliassi-Rad
Northeastern/Rutgers
A network is an eco-system

- Individuals have a mixture of roles in this eco-system
  - Roles = functions = positions
- Roles are defined in terms of structural behaviors
  - What is your connectivity pattern?
  - To what kinds of individuals are you connected?
Intuition: Types of neighbors matter

Node sizes indicate communication volume relative to the central node in each frame.
Finding roles in a network

**Input**

Node × Node Matrix

\( n \text{ dim space} \)
Finding roles in a network

Input

Node × Node Matrix

Recursive Feature Extraction

Node × Role Matrix

Role × Feature Matrix

n dim space

f dim space

[ReFex: Henderson et al., KDD 2011]
Finding roles in a network

**Input**

Node × Node Matrix

$n$ dim space

Recursive Feature Extraction

[ReFex: Henderson et al., KDD 2011]

Node × Feature Matrix

$f$ dim space

Role Extraction

[Rolx: Henderson et al., KDD 2012]

Node × Role Matrix

$r$ dim space

Role × Feature Matrix

**Output**

$n >> f >> r$
Finding roles in a network

Input

Node × Node Matrix

Recursive Feature Extraction

[ReFex: Henderson et al., KDD 2011]

Add guidance encoded as constraints on role assignments or role definitions

f dim space

Role Extraction + Guidance

[GLRD: Gilpin et al., KDD 2013]

Output

Node × Role Matrix

Role × Feature Matrix

r dim space

n >> f >> r
Big-data business-partnerships

- **Green**: equal opportunity bridges; big-data vendors
- **Red**: middle-men; general IT vendors.
- **Blue**: Strong affinity for big-data vendors; small vendors.
Big-data business-partnerships
Roles & communities are complementary

• Roles group nodes with similar structural properties

• Communities group nodes that are well-connected to each other
• **Green role:** equal opportunity bridges; big-data vendors.
• **Red role:** middle-men; general IT vendors.
• **Blue role:** Strong affinity for big-data vendors; small vendors.
Moving beyond simple networks

• Suppose you have a multi-relational networks
• Example: Congressional co-sponsorship data
No longer have an adjacency matrix

- We have a $\text{person} \times \text{person} \times \text{committee}$ tensor
- Entry at $(i, j, k)$ indicates how often congress-person $i$ and $j$ co-sponsored a bill that was sent to committee $k$ for a particular congressional committee

![Diagram of role discovery]

<table>
<thead>
<tr>
<th>Number of Roles</th>
<th>Accuracy</th>
<th>Accuracy (Var)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>67.3%</td>
<td>2.5</td>
</tr>
<tr>
<td>15</td>
<td>53.6%</td>
<td>3.9</td>
</tr>
<tr>
<td>20</td>
<td>43.1%</td>
<td>4.3</td>
</tr>
<tr>
<td>30</td>
<td>32.3%</td>
<td>5.2</td>
</tr>
</tbody>
</table>

Table 2: The interpretation of the best result is that we can re-identify the same person with a two thirds accuracy.

Table 3: The predictive accuracy of k-nearest neighbor algorithm for prediction to clearly show the benefits of our role representation. Table 3 shows experimental results for 10-fold cross validation to predict party (Democrat vs Republican) and discretized experience. We perform transfer learning. Transfer learning involves transferring in knowledge to solve a different problem. We use the simple k-nearest-neighbor algorithm for prediction to clearly show the benefits of our role representation. Table 3 shows experimental results for 10-fold cross validation to predict party (Democrat vs Republican) and discretized experience. We transfer mechanism by transferring in roles learnt from one congress (graph) to another. Our experiment involves applying our multi-relational role discovery method to find the roles used in earlier congresses by applying algorithm 1 but not solving for the roles themselves. If we do this for all congresses to create a richer role space/description.

![Graph with labels and roles]

http://mulan.sourceforge.net/datasets-mlc.html
Finding roles in a multi-relational network

- Multi-relational Role Discovery (MRD)
  - No orthogonality constraint on factors
  - Nonnegative Tucker decomposition
  - Alternating least squares

[MRD: Gilpin et al., under review]
Finding roles in a multi-relational network

- **Multi-relational Role Discovery (MRD)**
  - No orthogonality constraint on factors
  - Nonnegative Tucker decomposition
  - Alternating least squares

- The factor matrices are:
  - groups of features (*role definitions*)
  - groups of entities (*groups*)
  - groups of relations (*topics*)

- Tucker core

*MRD*: Gilpin et al., under review
## MRD Algorithm

**Algorithm 1** Multi-relational Role Discovery (MRD) using Alternating Least Squares Nonnegative Tucker decomposition.

1: Initialize $G, F, R$ and $H$ to any non-negative values
2: **while** Stop condition not met **do**
3: $G \leftarrow \text{argmin}_{G \geq 0} \| V_G - G H G (R \otimes F)^T \|_{Fro}$
4: Normalize the columns of $G$
5: $F \leftarrow \text{argmin}_{F \geq 0} \| V_F - F H F (R \otimes G)^T \|_{Fro}$
6: Normalize the columns of $F$
7: $R \leftarrow \text{argmin}_{R \geq 0} \| V_R - R H R (F \otimes G)^T \|_{Fro}$
8: Normalize the columns of $R$
9: $H \leftarrow \text{argmin}_{H \geq 0} \| \text{vec}(V) - (R \otimes F \otimes G) \text{vec}(H) \|_{Fro}$
10: **end while**
11: **return** $G, F, R, H$
Experiments

• Data from U.S. House of Representatives

• Bill co-sponsorship data from 1979 (the start of the 96th Congress) to 2009 (the end of the 110th Congress)

• 15 committees, for which there were legislation in each congress from 96th to 110th

• 110th Congress (from 2007-09)
  • 453 representatives & 10,613 bills
  • Average degree in aggregated graph = 8.37
  • Median value of average degree across committee co-sponsorship graphs = 0.48

<table>
<thead>
<tr>
<th>Committee</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sci &amp; Tech</td>
</tr>
<tr>
<td>Judiciary</td>
</tr>
<tr>
<td>Ways &amp; Means</td>
</tr>
<tr>
<td>VA</td>
</tr>
<tr>
<td>Small Business</td>
</tr>
<tr>
<td>Budget</td>
</tr>
<tr>
<td>Oversight &amp; Gov’t Reform</td>
</tr>
<tr>
<td>Agriculture</td>
</tr>
<tr>
<td>Appropriations</td>
</tr>
<tr>
<td>Rules</td>
</tr>
<tr>
<td>Natural Resources</td>
</tr>
<tr>
<td>Financial Services</td>
</tr>
<tr>
<td>Education &amp; Labor</td>
</tr>
<tr>
<td>Transportation &amp; Infrastructure</td>
</tr>
<tr>
<td>Energy &amp; Commerce</td>
</tr>
</tbody>
</table>
Model order selection

• Can do model order selection with Tucker
    • Automatic relevance determination (ARD)
    • A Bayesian approach that estimates the adequate degree of regularization
  • In these experiments, we set the model order to a $5 \times 5 \times 5$ core
Role definitions

[MRD: Gilpin et al., under review]
Role sense-making procedure

1. Run MRD to get the core and factor matrices: $V_1$, $H_1$, $G_1$, $R_1$, $F_1$.

2. Generate a new input tensor (nodes $\times$ relations $\times$ features), where the features are from a reference set of widely used and known features: $V_2$.

3. Use $V_2$, $H_1$, $G_1$, and $F_1$ to compute a new $R_2$ role definitions that make “sense” to a human.

Output: $R_2$, where roles are redefined in terms of a set of reference features each of which is normalized for comparison purposes.
Role sense-making in the 110\textsuperscript{th} Congress

- Role 3: Power brokers, high on every features
Role sense-making in the 110th Congress

- Role 1 & Role 4:
  - Both are path-y and on the periphery (high eccentricity values)
  - Both have very low degrees
  - But Role 4 nodes are more clique-y than Role 1 nodes (higher clust coeff) and less important (as measured by PageRank)
Role sense-making in the 110th Congress

- Role 2 & Role 5:
  - Both have high degrees and clust coeff
  - But Role 5 nodes have higher weight and higher PageRank
  ➞ Role 5 folks co-sponsor with the same people more often

![Graph showing property contributions for different roles.](image-url)
Relational topic definitions
Relational topics found

Topic 1
Ways & Means
Financial Services
Transportation & Infrastructure

Topic 2
Rules
Appropriations
Science & Technology

Topic 3
Oversight & Gov’t Reform
Education & Labor
Judiciary

Topic 4
Education & Labor
Natural Resources
VA

Topic 5
Agriculture
Science & Technology
Natural Resources
Topic 1: Ways & Means, Financial Services
Topic 2: Rules, Appropriations, S&T

Relational Topic 2

Science and Technology
Budget
Natural Resources
Judiciary
Oversight and Government Reform
Financial Services
Ways and Means
Agriculture
Education and Labor
Veterans’ Affairs
Appropriations
Transportation and Infrastructure
Small Business
Rules
Energy and Commerce
Topic 3: Oversight & Gov’t Reform, Education & Labor, Judiciary
Topic 4: Education & Labor, Natural Resources, VA
Topic 5: Agriculture, S&T, Natural Resources

Relational Topic 5

Science and Technology
Budget
Natural Resources
Judiciary
Oversight and Government Reform
Financial Services
Ways and Means
Agriculture
Education and Labor
Veterans’ Affairs
Appropriations
Transportation and Infrastructure
Small Business
Rules
Energy and and Commerce
Group definitions

[MRD: Gilpin et al., under review]
Groups of representatives

Group Members 1

- Millender–McDonald Juanita
- Obey David R.
- Tsongas Niki
- Speier Jackie
- Faleomavaega Eni F.H.
- Meehan Martin T.
- Edwards Donna F.
- Visclosky Peter J.
- Hoyer Steny H.
- Foster Bill
- Clyburn James E.
- Richardson Laura
- Becerra Xavier
- Pelosi Nancy
- Waters Maxine
- Velazquez Nydia M.
- Lantos Tom
- Childers Travis
- Cazayoux Donald J. Jr.
- Dicks Norman D.
Group 1 of representatives

<table>
<thead>
<tr>
<th>Name</th>
<th>Party</th>
<th>Exp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Millender-McDonald</td>
<td>D</td>
<td>11</td>
</tr>
<tr>
<td>Obey, David</td>
<td>D</td>
<td>38</td>
</tr>
<tr>
<td>Tsongas, Niki</td>
<td>D</td>
<td>0</td>
</tr>
<tr>
<td>Speier, Jackie</td>
<td>D</td>
<td>0</td>
</tr>
<tr>
<td>Faleomavaegag, Eni</td>
<td>D</td>
<td>18</td>
</tr>
<tr>
<td>Meehan, Martin</td>
<td>D</td>
<td>14</td>
</tr>
<tr>
<td>Edwards, Donna</td>
<td>D</td>
<td>0</td>
</tr>
<tr>
<td>Visclosky, Peter</td>
<td>D</td>
<td>22</td>
</tr>
<tr>
<td>Hoyer, Steny</td>
<td>D</td>
<td>26</td>
</tr>
<tr>
<td>Foster, Bill</td>
<td>D</td>
<td>0</td>
</tr>
</tbody>
</table>
More insights into Group 1

<table>
<thead>
<tr>
<th>Name</th>
<th>Party</th>
<th>Exp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Millender-McDonald</td>
<td>D</td>
<td>11</td>
</tr>
<tr>
<td>Obey, David</td>
<td>D</td>
<td>38</td>
</tr>
<tr>
<td>Tsongas, Niki</td>
<td>D</td>
<td>0</td>
</tr>
<tr>
<td>Speier, Jackie</td>
<td>D</td>
<td>0</td>
</tr>
<tr>
<td>Faleomavaega, Eni</td>
<td>D</td>
<td>18</td>
</tr>
<tr>
<td>Meehan, Martin</td>
<td>D</td>
<td>14</td>
</tr>
<tr>
<td>Edwards, Donna</td>
<td>D</td>
<td>0</td>
</tr>
<tr>
<td>Visclosky, Peter</td>
<td>D</td>
<td>22</td>
</tr>
<tr>
<td>Hoyer, Steny</td>
<td>D</td>
<td>26</td>
</tr>
<tr>
<td>Foster, Bill</td>
<td>D</td>
<td>0</td>
</tr>
</tbody>
</table>

Group 1
- Democrats; mostly not mid-career
- Active in oversight & gov’t reform
- On the periphery, but lots of triangles
### Group 2

<table>
<thead>
<tr>
<th>Name</th>
<th>Party</th>
<th>Exp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hensarling, Jeb</td>
<td>R</td>
<td>4</td>
</tr>
<tr>
<td>Boehner, John</td>
<td>R</td>
<td>16</td>
</tr>
<tr>
<td>Thornberry, Mac</td>
<td>R</td>
<td>12</td>
</tr>
<tr>
<td>Broun, Paul</td>
<td>R</td>
<td>0</td>
</tr>
<tr>
<td>Shadegg, John</td>
<td>R</td>
<td>12</td>
</tr>
<tr>
<td>Hastert, Dennis</td>
<td>R</td>
<td>8</td>
</tr>
<tr>
<td>Scalise, Steve</td>
<td>R</td>
<td>11</td>
</tr>
<tr>
<td>Latta, Robert</td>
<td>R</td>
<td>6</td>
</tr>
<tr>
<td>Flake, Jeff</td>
<td>R</td>
<td>6</td>
</tr>
<tr>
<td>McCreery, Jim</td>
<td>R</td>
<td>14</td>
</tr>
</tbody>
</table>

- Republicans
- Different topics
- Different roles
## Group 3

<table>
<thead>
<tr>
<th>Name</th>
<th>Party</th>
<th>Exp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cooper, Jim</td>
<td>D</td>
<td>16</td>
</tr>
<tr>
<td>Johnson, Henry</td>
<td>D</td>
<td>0</td>
</tr>
<tr>
<td>Ryan, Tim</td>
<td>D</td>
<td>4</td>
</tr>
<tr>
<td>DeGette, Diana</td>
<td>D</td>
<td>10</td>
</tr>
<tr>
<td>Engel, Eliot L.</td>
<td>D</td>
<td>14</td>
</tr>
<tr>
<td>Doggett, Lloyd</td>
<td>D</td>
<td>12</td>
</tr>
<tr>
<td>Pastor, Ed</td>
<td>D</td>
<td>16</td>
</tr>
<tr>
<td>Meek, Kendrick</td>
<td>D</td>
<td>4</td>
</tr>
<tr>
<td>Murphy, C.</td>
<td>D</td>
<td>0</td>
</tr>
<tr>
<td>Crowley, Joseph</td>
<td>D</td>
<td>8</td>
</tr>
</tbody>
</table>

### Group 3
- Democrats
- Same topic
- Different roles
Group 4

<table>
<thead>
<tr>
<th>Name</th>
<th>Party</th>
<th>Exp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hall, Ralph</td>
<td>R</td>
<td>16</td>
</tr>
<tr>
<td>Rodgers, Cathy</td>
<td>R</td>
<td>2</td>
</tr>
<tr>
<td>Myrick, Sue</td>
<td>R</td>
<td>12</td>
</tr>
<tr>
<td>Issa, Darrell</td>
<td>R</td>
<td>6</td>
</tr>
<tr>
<td>Drake, Thelma</td>
<td>R</td>
<td>2</td>
</tr>
<tr>
<td>Kuhl, Randy</td>
<td>R</td>
<td>2</td>
</tr>
<tr>
<td>Poe, Ted</td>
<td>R</td>
<td>2</td>
</tr>
<tr>
<td>Boozman, John</td>
<td>R</td>
<td>6</td>
</tr>
<tr>
<td>Conaway, Michael</td>
<td>R</td>
<td>2</td>
</tr>
<tr>
<td>Wamp, Zach</td>
<td>R</td>
<td>12</td>
</tr>
</tbody>
</table>

- Republicans
- Active in Agriculture
- High degree & very clique-y
**Group 5**

- Bipartisan
- Active in Agriculture
- Power brokers
Tucker core

MRD: Gilpin et al., under review
Interaction graph from the Tucker core
Measure properties on the interaction graph

<table>
<thead>
<tr>
<th>Property</th>
<th>Description</th>
<th>Computation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simplicity</td>
<td>To what extent are nodes connected to (role) similar types of nodes?</td>
<td>Average Node Degree</td>
</tr>
<tr>
<td>Sharing</td>
<td>How much can a group be separated into independent parts?</td>
<td>Mincut cost</td>
</tr>
<tr>
<td>Variability</td>
<td>How does the simplicity of nodes vary across the interaction graph?</td>
<td>Variance of node degree</td>
</tr>
<tr>
<td>Stability</td>
<td>How stable are the interactions between roles, groups, and topics?</td>
<td>Spectral gap</td>
</tr>
</tbody>
</table>
Cut cost of the interaction graphs from Tucker cores

More controversial bills passed
Role transfer (in this context)

• Roles extracted on one multi-relational network

• How well do the extracted roles transfer to another multi-relational network?
Role transfer on multi-relational Networks

Heatmap of fit quality $= 1$ – normalized reconstruction error

Roles Extracted on Congress $x$

Roles Transferred to Congress $y$
## Applications of role discovery

<table>
<thead>
<tr>
<th>Task</th>
<th>Use Case</th>
</tr>
</thead>
<tbody>
<tr>
<td>Role query</td>
<td>Identify individuals with similar behavior to a known target</td>
</tr>
<tr>
<td>Role outliers</td>
<td>Identify individuals with unusual behavior</td>
</tr>
<tr>
<td>Role dynamics</td>
<td>Identify unusual changes in behavior</td>
</tr>
<tr>
<td>Re-identification</td>
<td>Identify individuals in an anonymized network</td>
</tr>
<tr>
<td>Role transfer</td>
<td>Use knowledge of one network to make predictions in another</td>
</tr>
<tr>
<td>Network comparison</td>
<td>Determine network compatibility for knowledge transfer</td>
</tr>
<tr>
<td>Exploration in role space</td>
<td>Exploratory analysis of network data in the role space</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Why are roles effective?

• Encode complex behavior
• Map nodes into a useful lower dimensional space
• Generalize across networks
• Common language over a common alphabet

Funding from LLNL, NSF, IARPA, DARPA, and DTRA.
## Outline

<table>
<thead>
<tr>
<th>Time</th>
<th>Speaker</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00 - 9:50</td>
<td>David</td>
<td>Political inquiry, new science of politics, exemplary data</td>
</tr>
<tr>
<td>9:50 - 10:30</td>
<td>Oren</td>
<td>Exponential Random Graph Models</td>
</tr>
<tr>
<td>10:30 - 11:00</td>
<td></td>
<td>Coffee Break</td>
</tr>
<tr>
<td>11:00 - 11:20</td>
<td>Oren</td>
<td>Networks of political figures on Twitter</td>
</tr>
<tr>
<td>11:20 - 11:50</td>
<td>Tina</td>
<td>Roles in socio-political networks</td>
</tr>
<tr>
<td>11:50 - 12:00</td>
<td>David</td>
<td>Wrap-up &amp; questions</td>
</tr>
</tbody>
</table>
Wrap-up

- Tutorial website includes slides, resources (data & code)
  - [http://bit.ly/1Qs8blA](http://bit.ly/1Qs8blA)

- Seize the opportunity to create a new science of politics
THANKS!