WHEN COMPUTER SCIENCE MEETS SOCIAL SCIENCE

-- USER STANCE PREDICTION VIA INFORMATION NETWORK MINING

Yizhou Sun
Department of Computer Science
University of California, Los Angeles
yzsun@cs.ucla.edu

November 8, 2019

Outline

- Introduction
- Ideology Detection via Legislative Voting Network Mining
- Ideology Detection via Heterogeneous Types of Links
- User Stance Prediction via Joint Modeling of Text and Social Interactions
- The Co-Evolution Model for Social Network Evolving and Opinion Migration
- Summary

Motivating Examples

- What is the stance of Trump in the issue of military service?

- Whether a bill will be passed in congress?
- What is the average ideology of Twitter users in each state?
- By reading a news comment, can you infer the user's stance on that issue?
- Will a person's ideology migrate, and can you predict that?

GovTrack.us

Congress / Bills / H.R. 5404

H.R. 5404: Medical Device Guardians Act of 2016

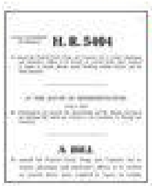
Introduced: **Jun 8, 2016**

Status: **Referred to Committee** on **Jun 8, 2016**
This bill was assigned to a congressional committee on June 8, 2016, which will consider it before possibly sending it on to the House or Senate as a whole.

Sponsor:

Michael Fitzpatrick
Representative for Pennsylvania's 8th congressional district
Republican

Text:



Read Text »
Last Updated: Jun 8, 2016
Length: 3 pages

Prognosis: 4% chance of being enacted (details)

TRACK THIS BILL

Call Congress

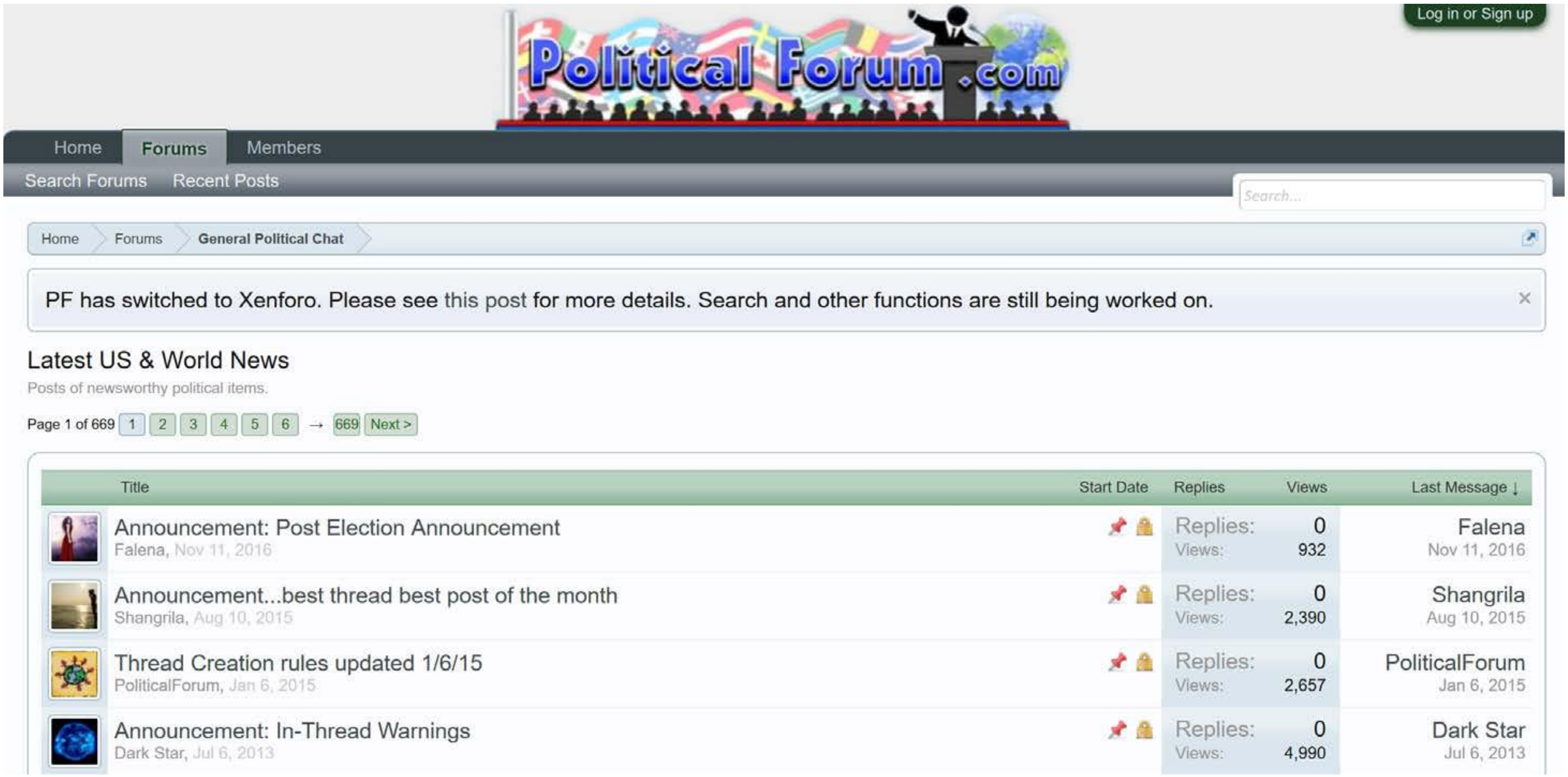
1

Twitter

TRUMP
MAKE AMERICA GREAT AGAIN!
TWEETS 27.5K
FOLLOWING 42
FOLLOWERS 3.75M
FAVORITES 56
Tweets
Tweets & replies
Photos & videos
Who to
Donald J. Trump
@realDonaldTrump
The official Twitter profile for Donald Trump donaldjtrump.com youtube.com/DonaldTrump facebook.com/DonaldTrump
New York, NY
donaldjtrump.com
Joined March 2009
Tweet to Donald J. Trump
Donald J. Trump @realDonaldTrump · 34m
The polls have been really amazing--we are all tired of incompetent politicians and bad deals! newsmax.com/Newsfront/fox-...
477
1K
Donald J. Trump @realDonaldTrump · 1h
I really enjoyed being at the Iowa State Fair. The crowds, love and enthusiasm is something I will never forget.
614
1.8K

News Comments

Online Forums



Outline

- Introduction
- Ideology Detection via Legislative Voting Network Mining
- Ideology Detection via Heterogeneous Types of Links
- User Stance Prediction via Joint Modeling of Text and Social Interactions
- The Co-Evolution Model for Social Network Evolving and Opinion Migration
- Summary

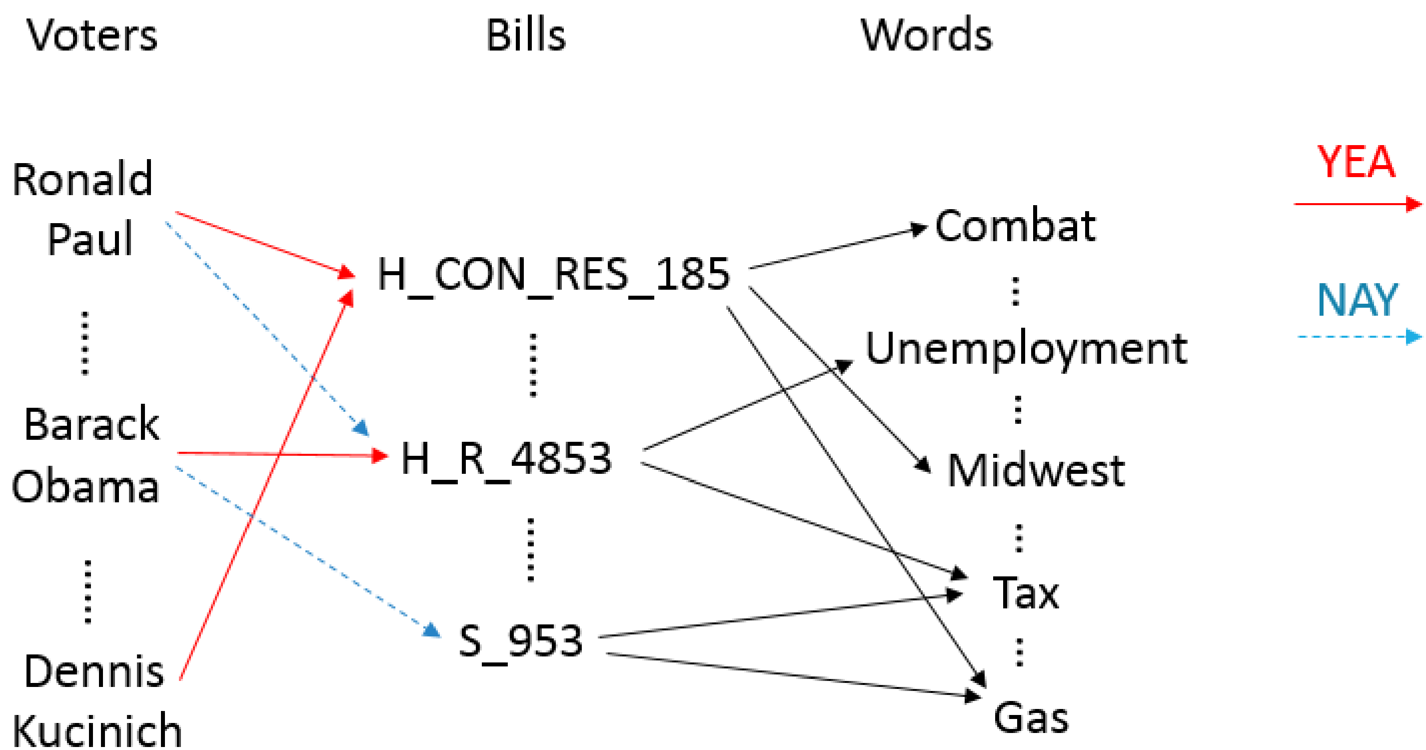
- Topic-Factorized Ideal Point Estimation Model for Legislative Voting Network (Gu et al., KDD'14)

Background of Congress Roll Call Data

Federal Legislation (bill) $\xrightarrow[\text{Senate}]{\text{The House}}$ Law

	Bill 1	Bill 2	……					
Ronald Paul	✓	✓	✗	✓	✗	✓	✓	
Barack Obama		✓	✓	✗	✓	✓	✓	✗
⋮								

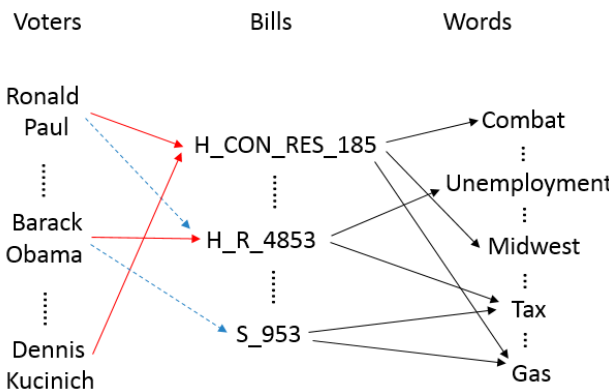
Legislative Voting Network

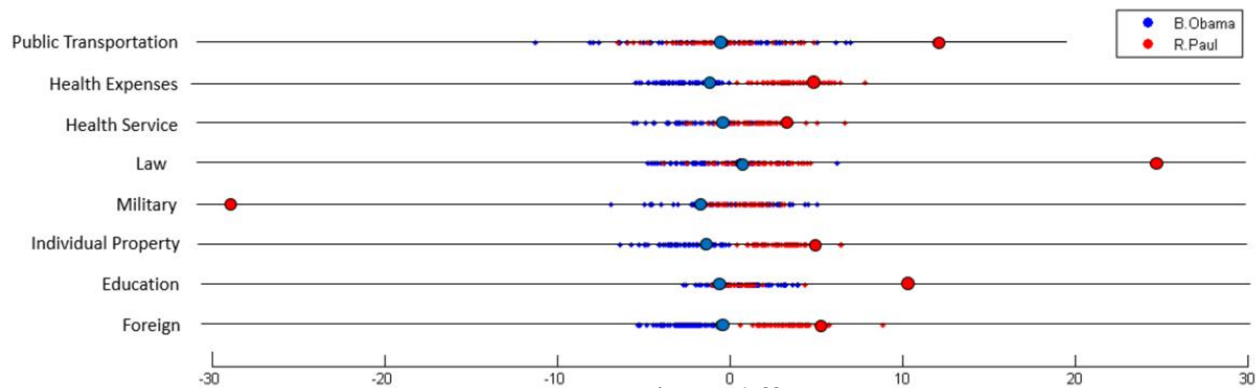


Problem Definition

Input:
Legislative Network

Output:
$\boldsymbol{x}_u$: Ideal Points for Politician u
$\boldsymbol{a}_d$: Ideal Points for Bill d





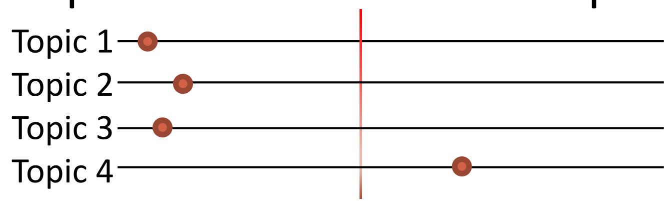
$\boldsymbol{x}_u$'s on different topics

Existing Work

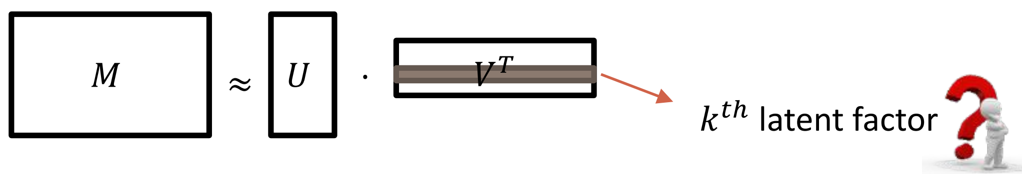
- 1-dimensional ideal point model (Poole and Rosenthal, 1985; Gerrish and Blei, 2011)
- High-dimensional ideal point model (Poole and Rosenthal, 1997)
- Issue-adjusted ideal point model (Gerrish and Blei, 2012)

Motivation

- Voters have different positions on different topics.



- Traditional matrix factorization method cannot give the meanings for each dimension.



- **Topics of bills can influence politician's voting, and the voting behavior can better guide the topics of bills as well.**

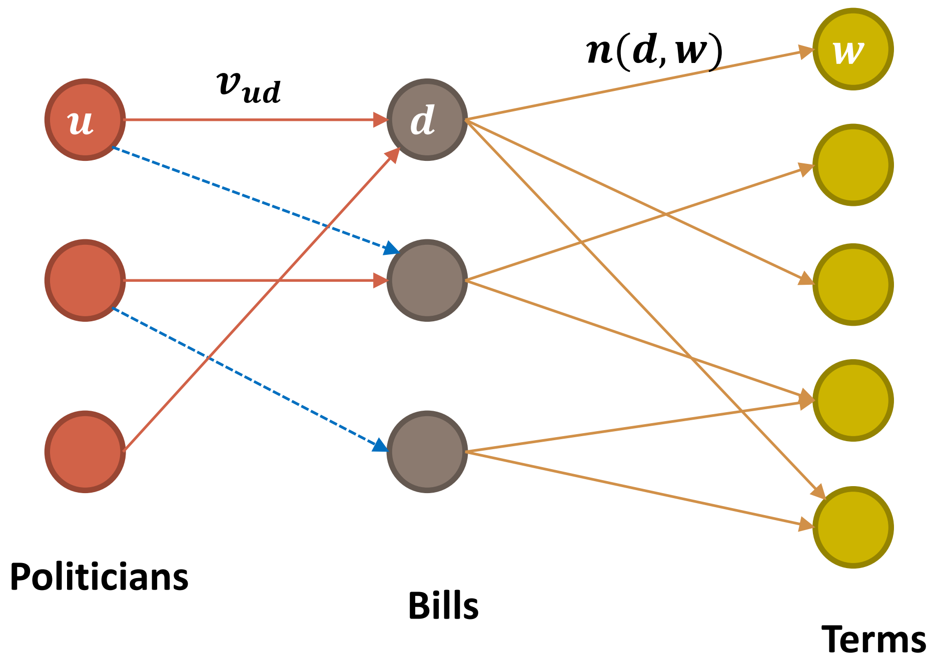
Topic Model:

- Health
- Public Transport
- …

Voting-guided Topic Model:

- Health Service
- Health Expenses
- Public Transport
- …

Topic-Factorized IPM



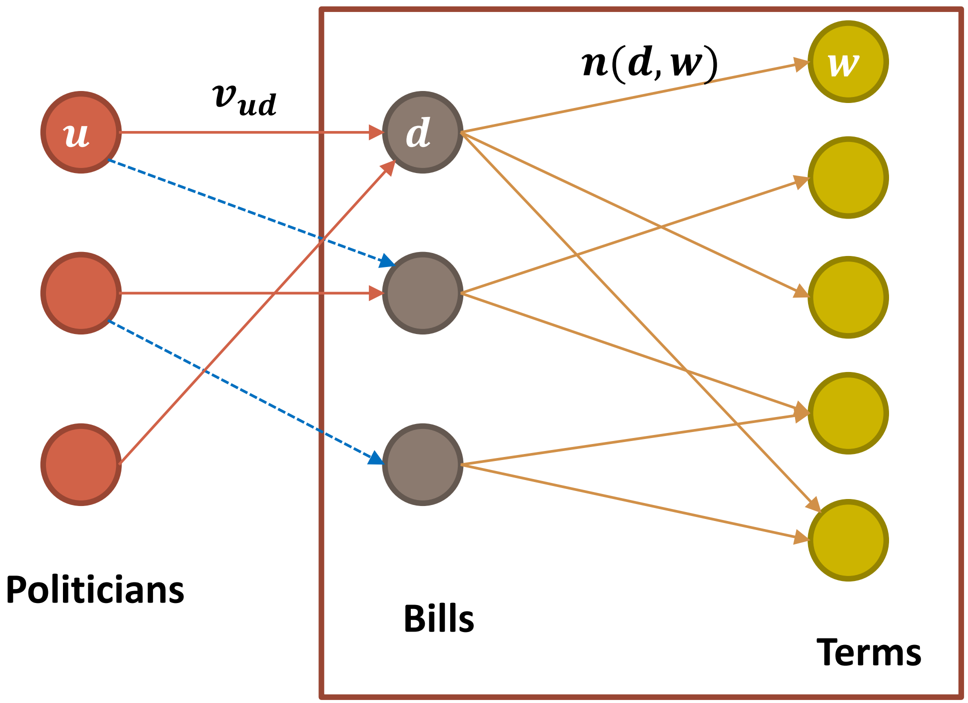
Heterogeneous Voting Network

Entities:
- Politicians
- Bills
- Terms

Links:
- (P, B)
- (B, T)

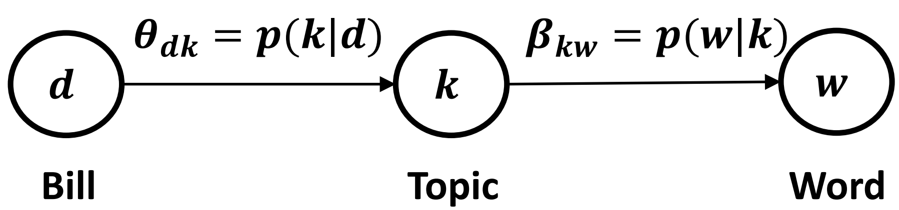
Parameters to maximize the likelihood of generating two types of links:
- Ideal points for politicians
- Ideal points for bills
- Topic models

Text Part



Text Part

- We model the probability of each word in each document as a mixture of categorical distributions, as in PLSA (Hofmann, 1999) and LDA (Blei et al., 2003)

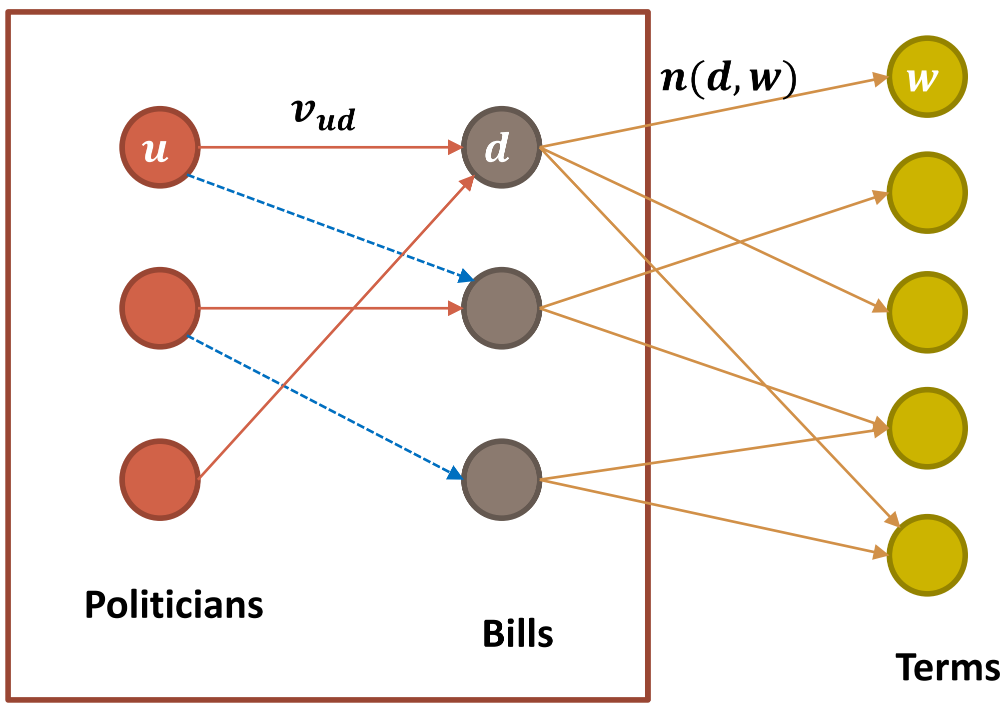


$$\boldsymbol{w}_d = \left(n(d,1), n(d,2), \ldots, n(d,N_w)\right)$$

$$p(\boldsymbol{w}_d|\boldsymbol{\theta},\boldsymbol{\beta}) \propto \prod_w \left(\sum_k \theta_{dk}\beta_{kw}\right)^{n(d,w)}$$

$$p(\boldsymbol{W}|\boldsymbol{\theta},\boldsymbol{\beta}) \propto \prod_d \prod_w \left(\sum_k \theta_{dk}\beta_{kw}\right)^{n(d,w)}$$

Voting Part



Intuitions:

- The more similar of the ideal points of *u* and *d*, the higher probability of “YEA” link
- The higher portion a bill belongs to topic *k*, the higher weight of ideal points on topic *k*

Voting Part

Ideology embedding of Voter u	x_u	x_{u1}	x_{u2}			x_{uk}		x_{uK}	$x_{uk} \in R$
Ideology embedding of Bill d	a_d	a_{d1}	a_{d2}			a_{dk}		a_{dK}	$a_{dk} \in R$
Topic Distribution of Bill d	θ_d	θ_{d1}	θ_{d2}			θ_{dk}		θ_{dK}	

YEA

$$p(v_{ud} = 1) = \sigma(\sum_k \theta_{dk} x_{uk} a_{dk} + b_d)$$

NAY

$$p(v_{ud} = -1) = 1 - \sigma(\sum_k \theta_{dk} x_{uk} a_{dk} + b_d)$$

$$p(V|\theta, X, A, b) = \prod_{(u,d):v_{ud} \neq 0} (p(v_{ud} = 1)^{\overset{I_{\{v_{ud}=1\}}}{\boxed{\frac{1+v_{ud}}{2}}}} p(v_{ud} = -1)^{\overset{I_{\{v_{ud}=-1\}}}{\boxed{\frac{1-v_{ud}}{2}}}})$$

Grounded in social choice theory: Choose YEA or NAY?

Combining Two Parts Together

- The final objective function is a linear combination of the two average log-likelihood functions over the *word links* and *voting links*.

$$J(\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{X},\boldsymbol{A},\boldsymbol{b}) = (1-\lambda)\boxed{\frac{\sum_{d,w} n(d,w)\log(\sum_k \theta_{dk}\beta_{kw})}{N_F}} + \lambda\boxed{\frac{\sum_{(u,d):v_{ud}\neq 0}(\frac{1+v_{ud}}{2}\log p(v_{ud}=1) + \frac{1-v_{ud}}{2}\log p(v_{ud}=-1))}{N_V}}$$

s.t.

$$0 \le \theta_{dk} \le 1, \quad \sum_k \theta_{dk} = 1 \qquad \text{and} \qquad 0 \le \beta_{kw} \le 1, \quad \sum_w \beta_{kw} = 1$$

- We also add an l_2 regularization term to A and X to reduce over-fitting.

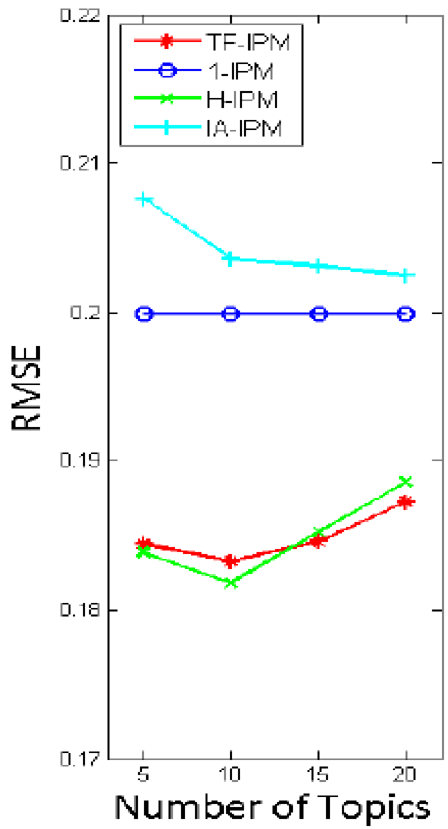
$$J(\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{X},\boldsymbol{A},\boldsymbol{b}) = (1-\lambda)\frac{\sum_{d,w} n(d,w)\log(\sum_k \theta_{dk}\beta_{kw})}{N_F} + \lambda\frac{\sum_{(u,d):v_{ud}\neq 0}(\frac{1+v_{ud}}{2}\log p(v_{ud}=1) + \frac{1-v_{ud}}{2}\log p(v_{ud}=-1))}{N_V}$$
$$\boxed{-\frac{1}{2\sigma^2}(\sum_u ||\boldsymbol{x}_{\boldsymbol{u}}||_2^2 + \sum_d ||\boldsymbol{a}_d||_2^2)}$$

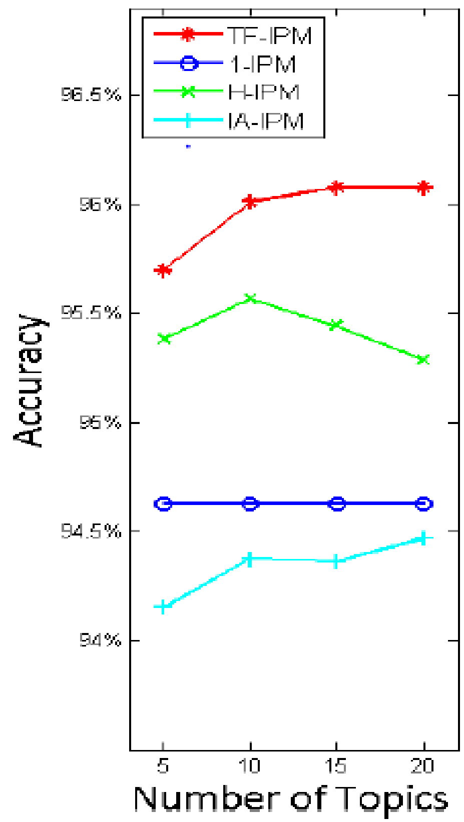
Data Description

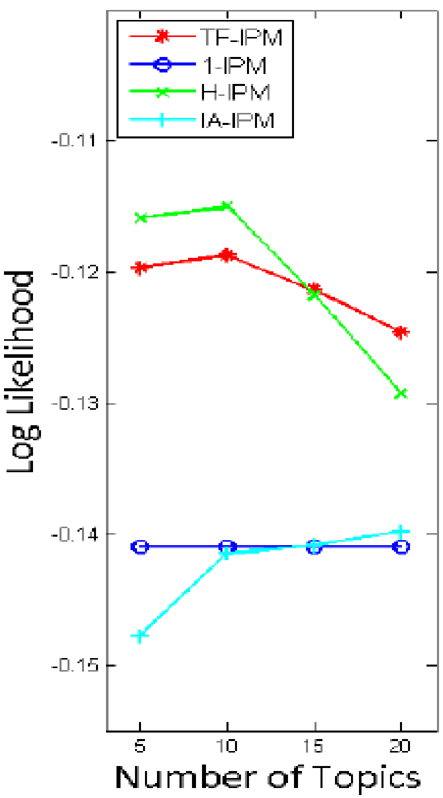
- Dataset:
  - U.S. House and Senate roll call data in the years between 1990 and 2013.*
    - 1,540 legislators
    - 7,162 bills
    - 2,780,453 votes (80% are "YEA")
  - Keep the latest version of a bill if there are multiple versions.
  - Randomly select 90% of the votes as training and 10% as testing.

* Downloaded from http://thomas.loc.gov/home/rollcallvotes.html

Experimental Results: Voting Result Prediction

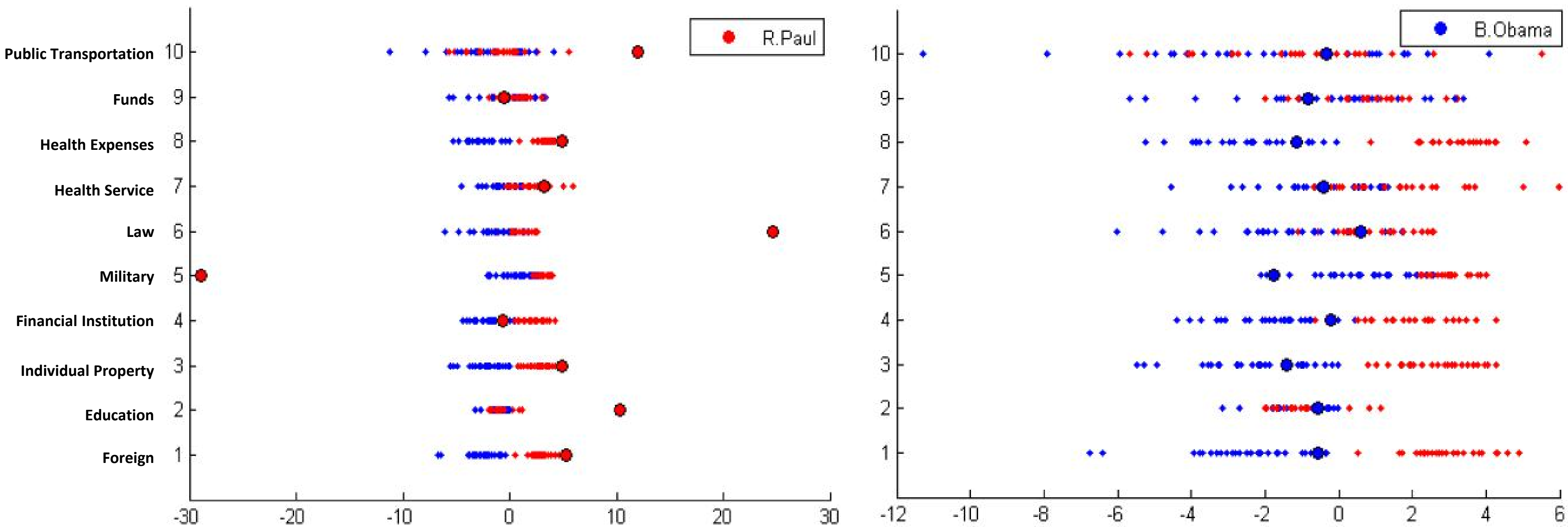






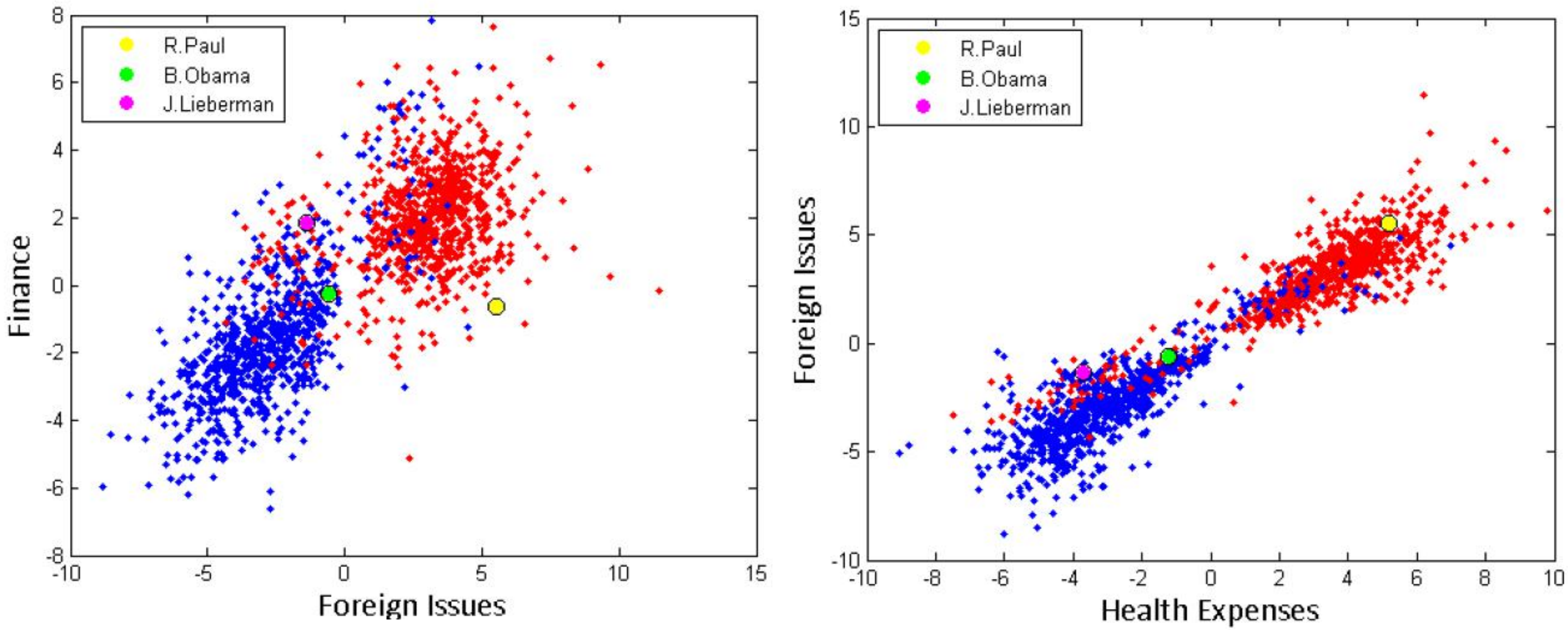
Case Studies

- Ronald Paul (R), Barack Obama (D)



Case Studies

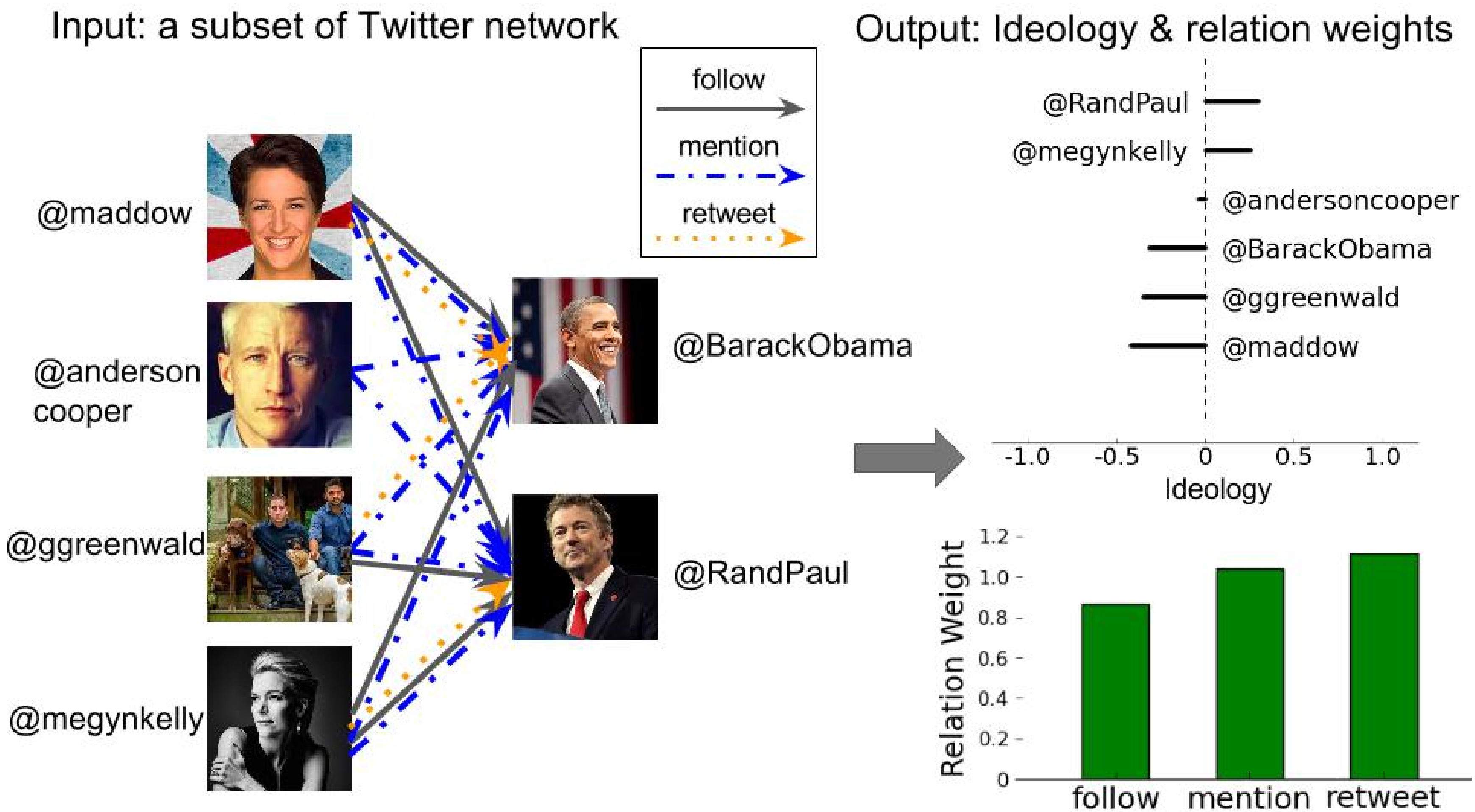
- Scatter plots over selected dimensions: (Republican, Democrat)



Outline

- Introduction
- Ideology Detection via Legislative Voting Network Mining
- Ideology Detection via Heterogeneous Types of Links
- User Stance Prediction via Joint Modeling of Text and Social Interactions
- The Co-Evolution Model for Social Network Evolving and Opinion Migration
- Summary

Ideology Detection for Twitter Users [Gu et al., SBP-BRiMS 2017]



Challenges

- How to model behavior links using ideology?
  - A link generative model with ideology as latent variable
  - Another social choice problem: to follow or not?

- How to combine different types of behaviors?
  - Heterogeneous network mining

Ideology Detection via Single Type of Links

- Assumptions
  - People tend to *follow, mention, or retweet* others who share similar ideology score
  - One's true ideology score is different from one's public image ideology score

Model: Take *Follow* Link Type as an Example

- Consider a follow link: $i \rightarrow j$
  - $\boldsymbol{p}_i$: true ideology of user i
  - $\boldsymbol{q}_j$: image ideology of user j
  - b_j: bias term of user j, indicating the popularity of user j

  - The probability of the link:
    - $p(i \rightarrow j) = \sigma(\boldsymbol{p}_i' \boldsymbol{q}_j + b_j)$ (σ is the sigmoid function)

  - Connection to utility function:
    - The utility function of user i choose to follow j is higher than the one of user i choose to not follow j.

Objective Function

- Maximize the likelihood of observed positive links and sample negative links

$$l(G) = \sum_{(i,j)\in S_+} e_{ij}^{+} \cdot \log \sigma_{ij} + \sum_{(i,j)\in S_-} e_{ij}^{-} \cdot \log(1 - \sigma_{ij})$$

Ideology Detection via Multiple Types of Links

- Consider a link of type r: $i \xrightarrow{r} j$
  - $\boldsymbol{p}_i$: true ideology of user i
  - $\boldsymbol{q}_j^{(r)}$: image ideology of user j for link type r
  - $b_j^{(r)}$: bias term of user j for link type r, indicating the popularity of user j in link type r

  - The probability of the link:
    - $p\left(i \xrightarrow{r} j\right) = \sigma(\boldsymbol{p}_i' \boldsymbol{q}_j^{(r)} + b_j^{(r)})$ (σ is the sigmoid function)

Objective Function

- Maximize the weighted sum of the likelihood function for each link type

$$l(\boldsymbol{G}|\boldsymbol{P},\boldsymbol{Q},\boldsymbol{B})$$
$$= \sum_{r=1}^{R} \boxed{w_r} \cdot \frac{\sum_{(i,j)\in S_+^{(r)}} e_{r,ij}^{+} \log \sigma_{r,ij} + \sum_{(i,j)\in S_-^{(r)}} e_{r,ij}^{-} \log(1-\sigma_{r,ij})}{\sum_{(i,j)\in S_+^{(r)}} e_{r,ij}^{+} + \sum_{(i,j)\in S_-^{(r)}} e_{r,ij}^{-}}$$

Importance of r_{th} link type

Constraints on w_r: $w_r > 0$ *and* $\boxed{\prod_r w_r = 1}$

Geometric mean =1

Data Collection

- Step 1: Seed users collection
  - Manually collect 487 politicians from Twitter
- Step 2: expand from seeds
  - 5000 followers and followees for each seed
- Step 3: filtering
  - User set I (heavily political related): users with more than 20 followers or followees in the seed set
  - User set II (less political related): users with 3-5 followers or followees in the seed set, random sample 10,000

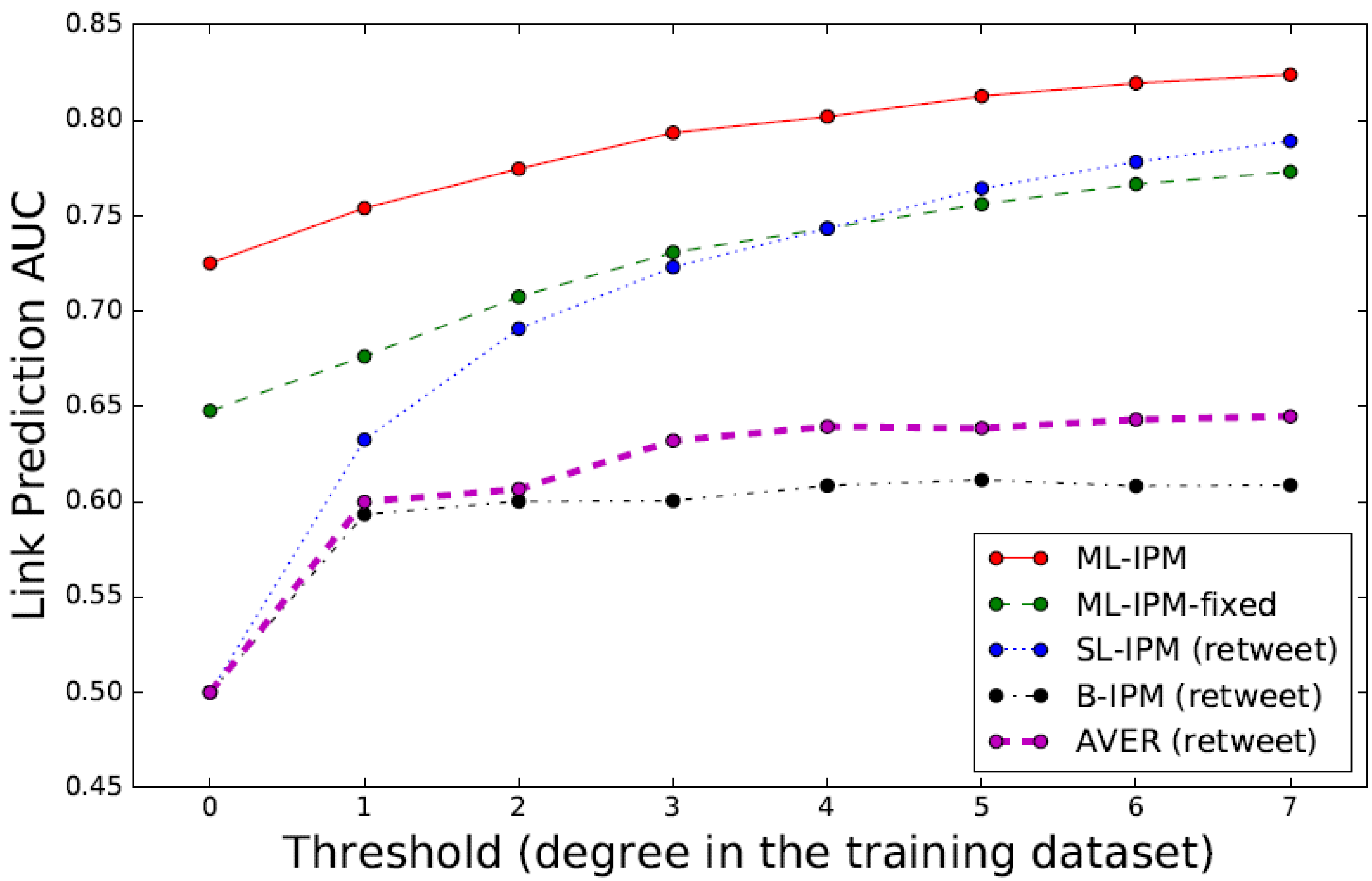
Relation	*follow*	*mention*	*retweet*
Number of links	1,764,956	2,395,813	718,124
Total number of users	46,477	34,775	30,990

Results on Ranking and Classification

Method	Ranking Accuracy	Classification AUC
AVER (*follow*)	0.427	0.523
AVER (*mention*)	0.446	0.558
AVER (*retweet*)	0.474	0.587
B-IPM (*follow*)	0.443 ± 0.102	0.868 ± 0.021
B-IPM (*mention*)	0.433 ± 0.183	0.558 ± 0.064
B-IPM (*retweet*)	0.501 ± 0.127	0.561 ± 0.066
SL-IPM (*follow*)	0.626 ± 0.011	0.953 ± 0.015
SL-IPM (*mention*)	0.623 ± 0.027	0.951 ± 0.018
SL-IPM (*retweet*)	0.637 ± 0.005	0.958 ± 0.005
ML-IPM-fixed	0.655 ± 0.008	0.930 ± 0.035
ML-IPM	$\mathbf{0.663 \pm 0.007}$	$\mathbf{0.986 \pm 0.013}$

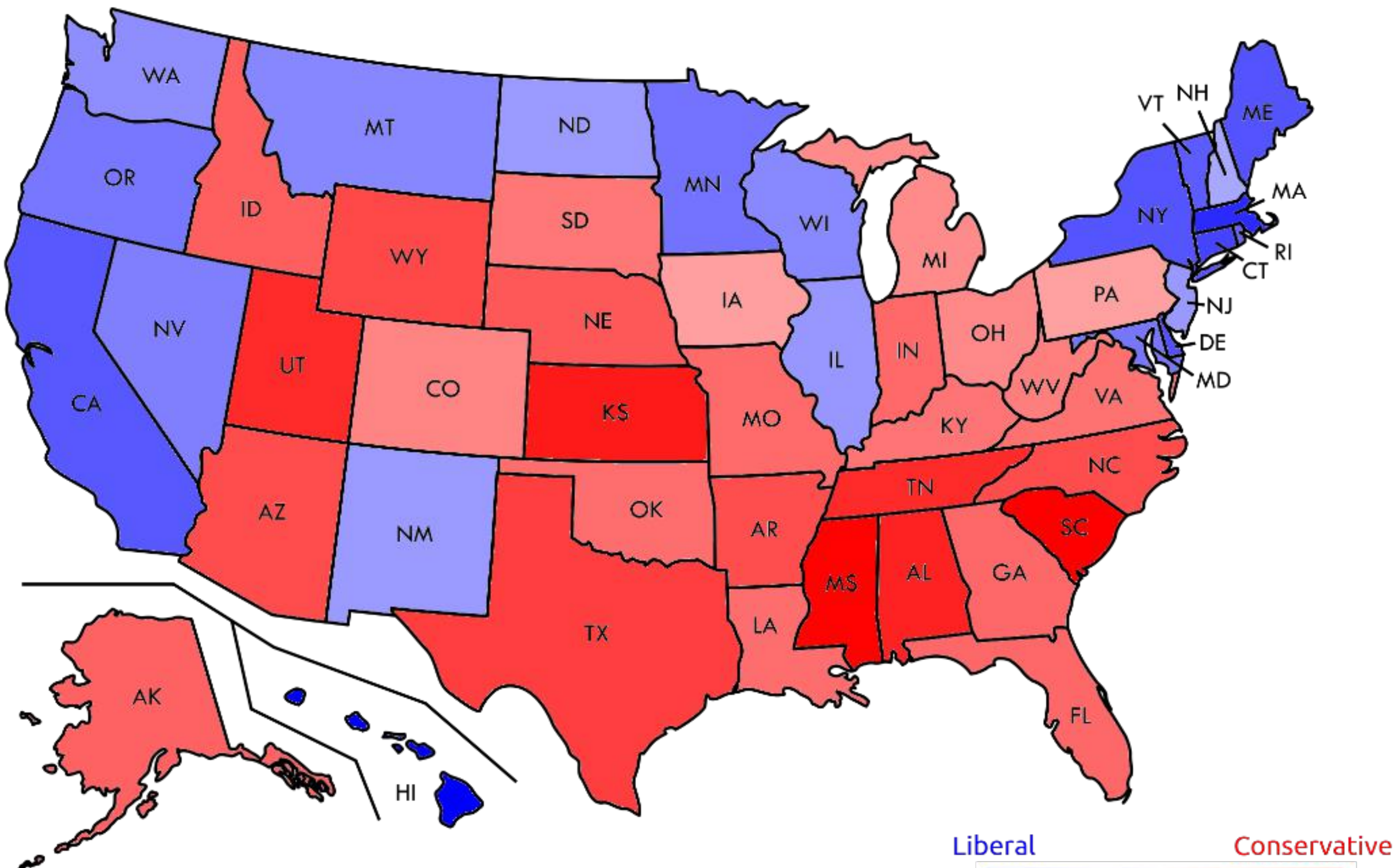
Retweet Link Prediction

- Using ideology to predict retweet link



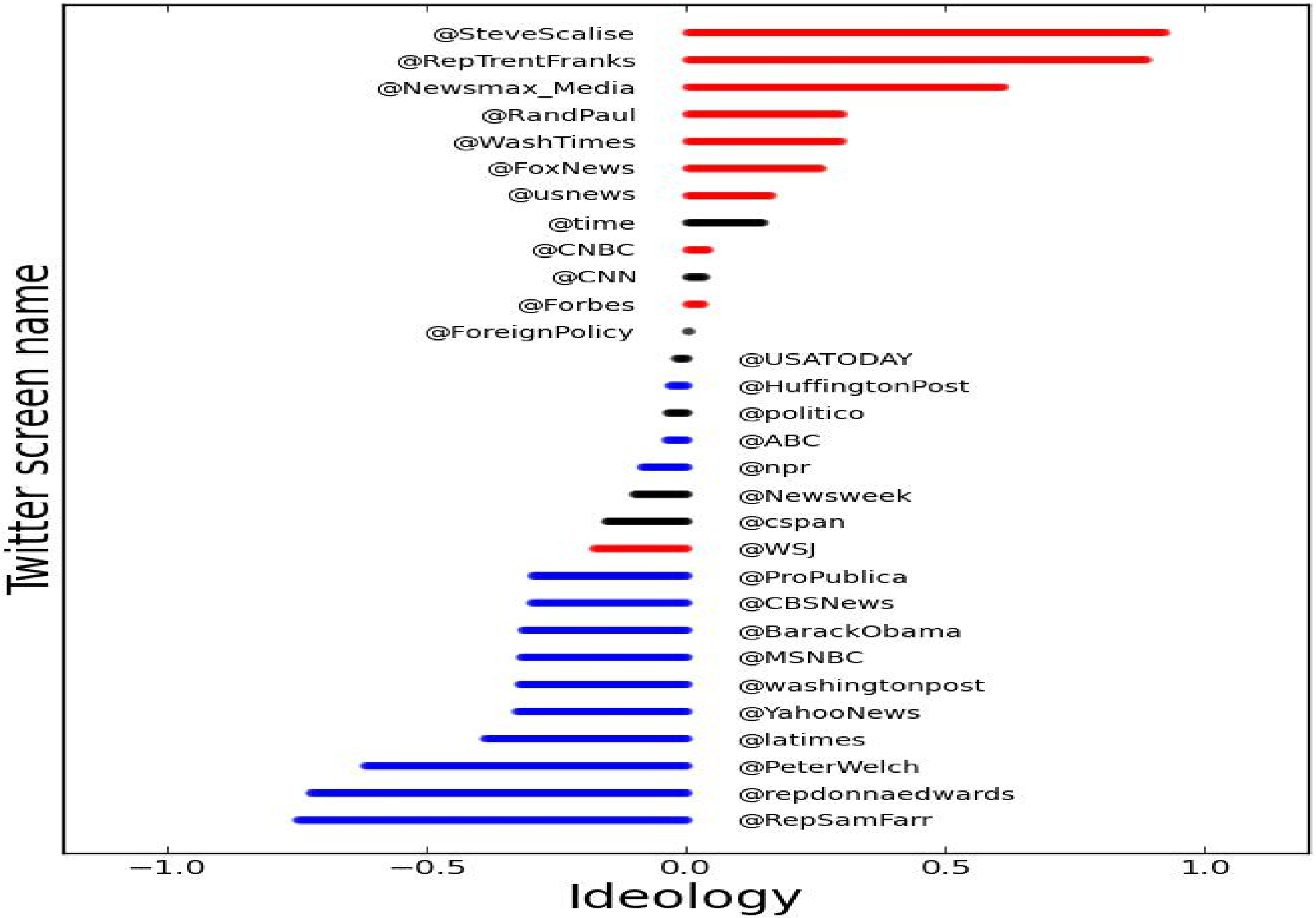
Case Study 1

- The ideology map of U.S.

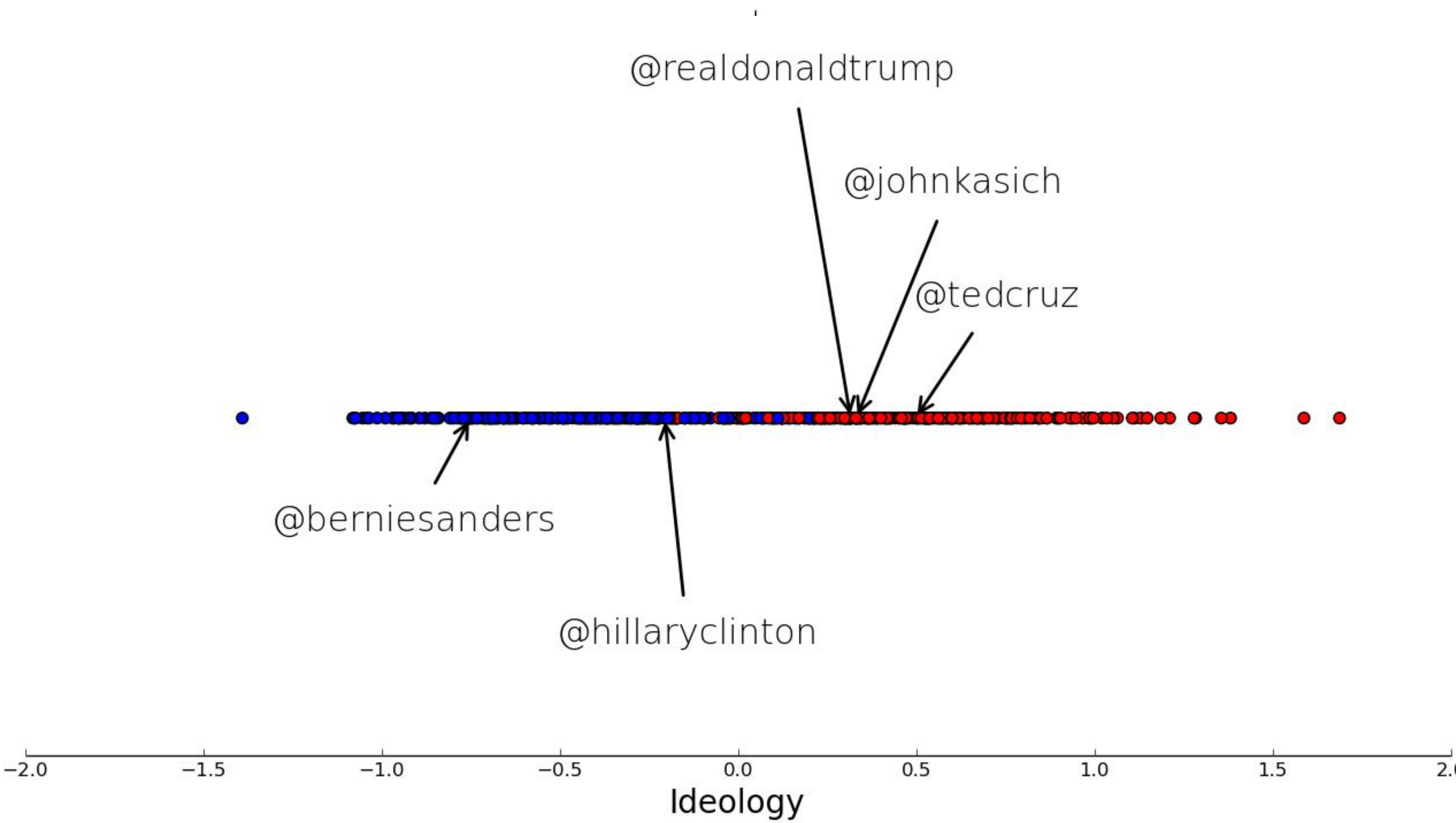


Case Study 2

- Ideology for selected accounts



Ideology of Presidential Candidates



Outline

- Introduction
- Ideology Detection via Legislative Voting Network Mining
- Ideology Detection via Heterogeneous Types of Links
- User Stance Prediction via Joint Modeling of Text and Social Interactions
- The Co-Evolution Model for Social Network Evolving and Opinion Migration
- Summary

Weakly-Guided User Stance Prediction via Joint Modeling of Content and Social Interaction (Dong et al., CIKM'17)

User1: Why don't you people stay in your bedrooms and quit flaunting your illicit pro **gay sex** sin agenda in public?

User2: I didn't know your church had its own... it is about living our lives with the person we love with full rights and benefits **under the law** … value of your union except in your own head.

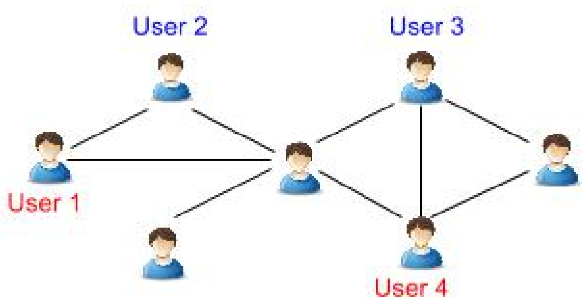
User3: It has been proven a dozen times over on … Gay Couples are denied the right for p.rotection **under the law** ... Denying **same sex couples** the right to marry is creating **second class** citizens ... Do your homework!!!!

User4: Yes, marriage IS a civil right ... But the Loving case was about a **man and woman** getting together, not same-gender people. Of course the … that doesn't mean what gay people are doing is a civil right.

(a) User discussions

Stance1	Stance2
under the law	**gay sex**
same sex couples	**word of god**
my son	***sodom and gomorrah***
gays and lesbians	my friend
quite frankly	gay pride
inter racial marriage	**man and woman**
partner and i	president bush
second class	**redefine marriage**
legitimate state interest	**deviant sex**
equal marriage	these forums

(c) Word distributions



(b) User interactions

Output

(d) User stances

Intuitions to the Solution

- With different stances, words are used differently even for the same issue
  - E.g., “same-sex marriage” vs. “gay marriage”

- User interaction will contribute to the learn the right contrasting view points for each issue
  - E.g., a reply indicating “disagree” implies the opposite signs of the related two users

Case Study from CNN News Comments

Bowe Bergdahl		Gaza Israel	
Stance 1	Stance 2	Stance 1	Stance 2
republicans	**obama**	hamas	**jew**
gop	**liberal**	**muslim**	**jews**
the gop	deserter	**free palestine from arad terror**	netanyahu
conservatives	**liberals**	egypt	**isreal**
allegedly	a deserter	yawn	israeli
Exu **republican**	arabs	hitler	**conservatives**
reagan	traitor	hamass	aipac
conservative	he deserted	syria	part of this genocide
fox news	obama is	hamas_is	**cut all aid to israel**
right wing	susan rice	**muslims**	zionists

Immigrant		MH17	
Stance 1	Stance 2	Stance 1	Stance 2
republicans	**obama**	**putin**	**usa**
the gop	**liberals**	**russian**	**kiev**
gop	**democrats**	russians	ukrainian
boehner	**illegals**	the russians	iraq
republican	liberal	**kremlin**	**americans**
conservatives	the illegals	russia	american
congress	illegal aliens	vodka	**cia**
perry	obama is	comrade	poroshenko
the republicans	**dems**	**russian troll**	com watch
conservative	citizens	huh	youtube

Outline

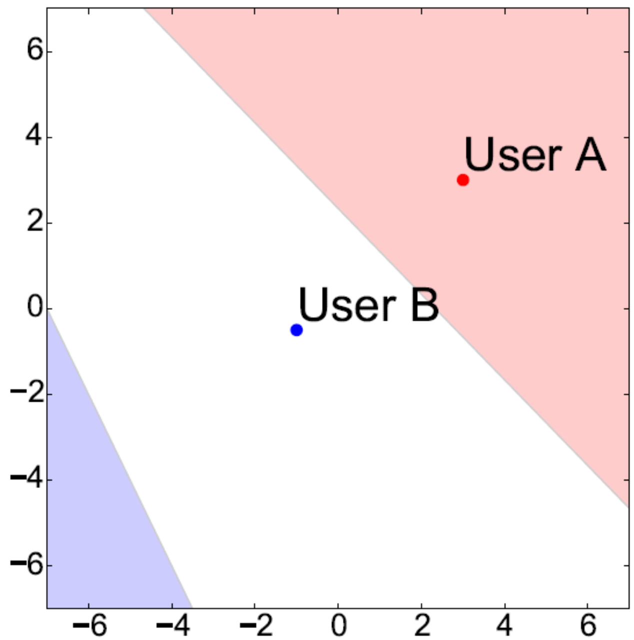
- Introduction
- Ideology Detection via Legislative Voting Network Mining
- Ideology Detection via Heterogeneous Types of Links
- User Stance Prediction via Joint Modeling of Text and Social Interactions
- The Co-Evolution Model for Social Network Evolving and Opinion Migration
- Summary

- The Co-Evolution Model for Social Network Evolving and Opinion Migration (Gu et al., KDD'17)

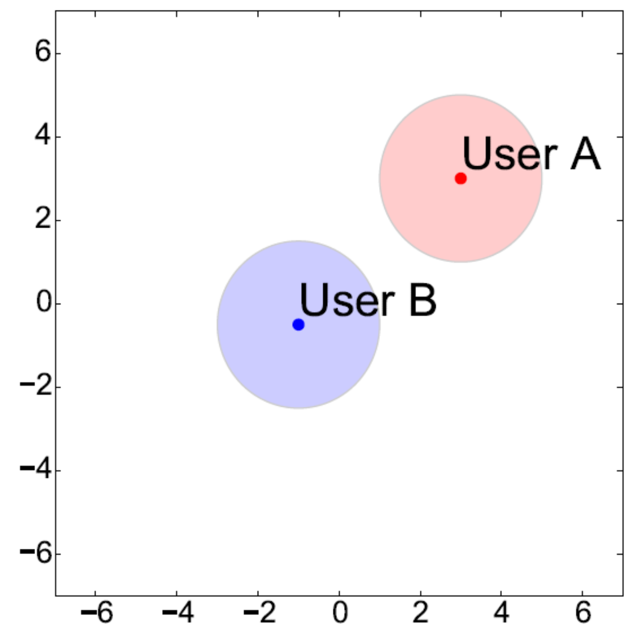
Network Generation

- People with similar ideology tends to link together

- How to define similarity?



Dot product-based similarity



Euclidean distance-based similarity

Gravity-based Score Function

- Observations
  - Opinion leader have more connections
    - Model opinion leaders explicitly: b_i $for\ actor\ i$
  - Opinion leaders tend to have a non-extreme ideology
    - Choose Euclidean distance-based score
- The score function: extending Gaussian Kernel

$$p_{ij} = \exp\left(-\frac{1}{\epsilon^2} \cdot \frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{b_i \cdot b_j}\right)$$

- Link Generation:
  - $G_{ij} = 1, if\ p_{ij} > d$, where d is the system parameter controlling how easy to form a link

Opinion Migration

- Actors are influenced by their neighbors

- What are influenced?
  - Position or moving direction?

Opinion Propagation Model

- Moving directions determine the position
  - Continuous form:

$$\frac{d}{dt}\boldsymbol{x}_n(t) = v \cdot (\cos\theta_n(t),\ \sin\theta_n(t))$$

  - Discrete from:

$$\boldsymbol{x}_n^{\langle t+1\rangle} = \boldsymbol{x}_n^{\langle t\rangle} + v \cdot (\cos\theta_n^{\langle t\rangle},\ \sin\theta_n^{\langle t\rangle})$$

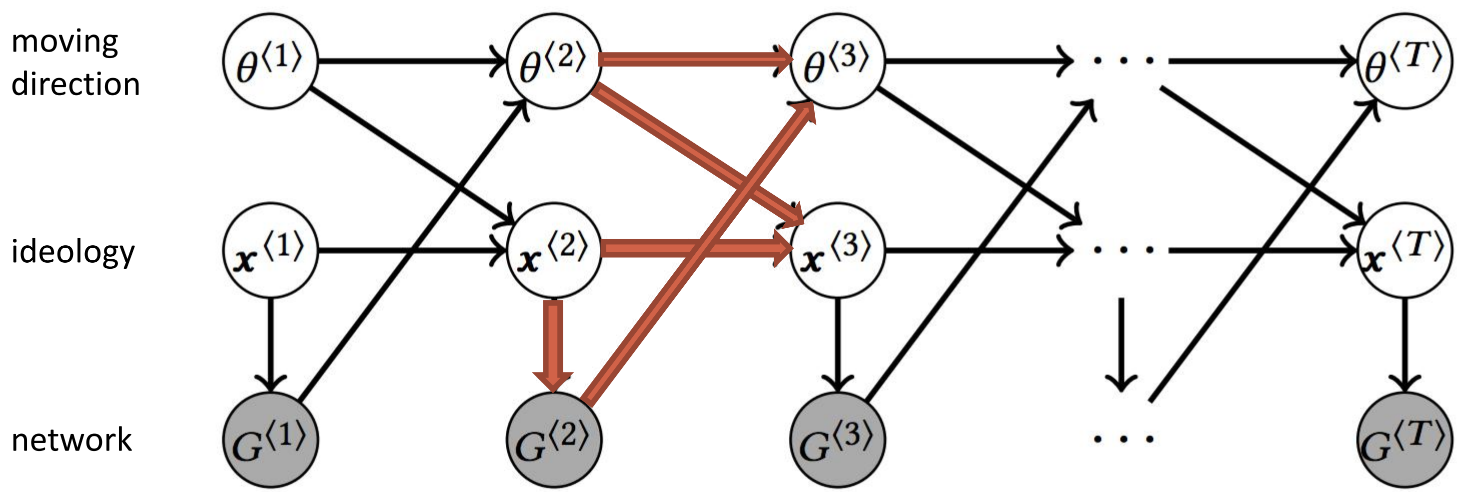
- Moving directions are influence by social neighbors

$$\theta_n^{\langle t+1\rangle} \sim \mathcal{N}(\langle\theta_n^{\langle t\rangle}\rangle, \sigma^2)$$

$\boldsymbol{\sigma^2}$: ***system level parameter controling the noise***

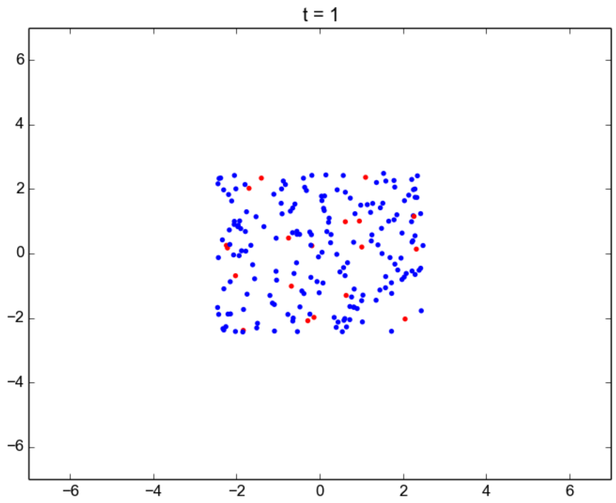
Putting it together

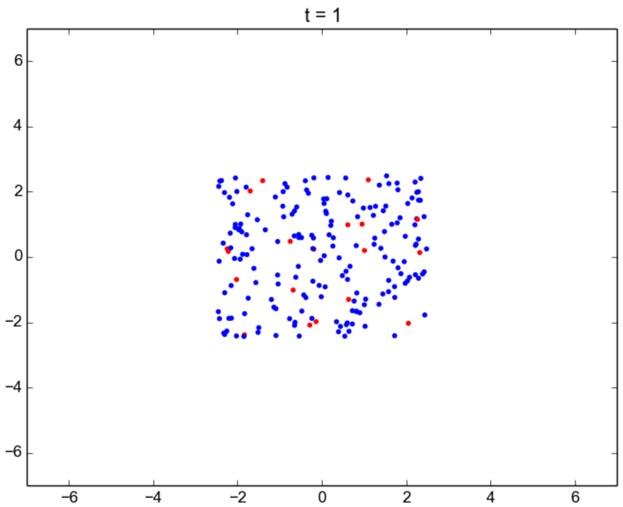
- The co-evolution model



Simulation

- System level parameters:
  - Actor popularity: b
  - Network sparsity: d
  - Noise in propagation: σ^2





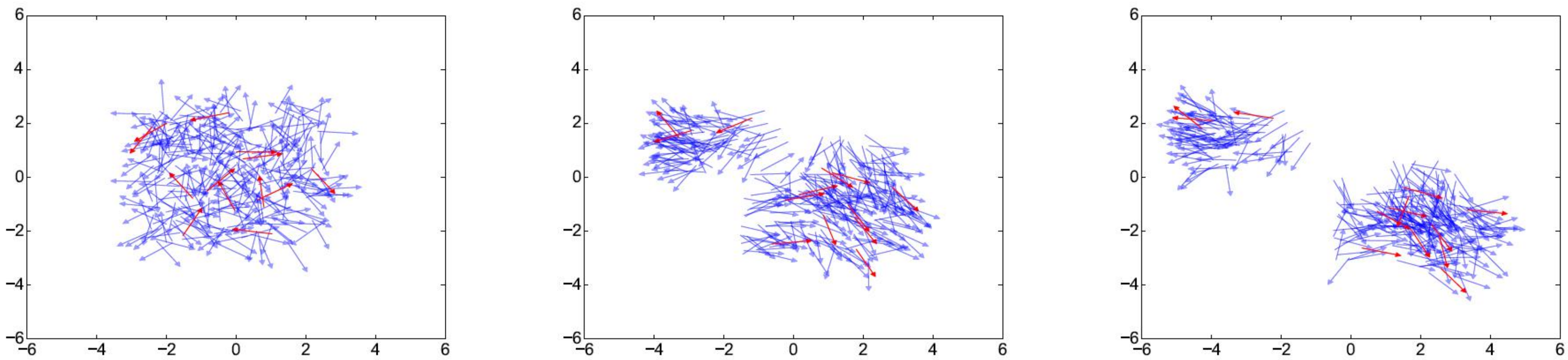
Discovery

- System level parameters can control the behavior of evolution

Noise level σ	Sparsity parameter d	Result
Small	Large (fewer neighbors)	Opinion divergence (emergence of clusters)
Small	Small (more neighbors)	Opinion convergence
Large	-	Random

Small noise, fewer friends

- Communities appear, opinion divergence

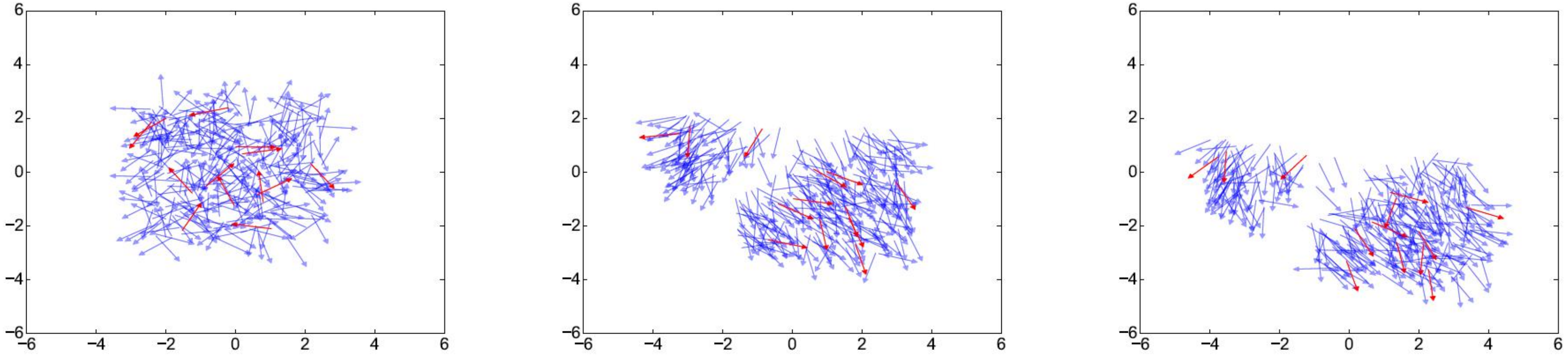


(a) $t = 0$ (init) (b) $t = 40$ (c) $t = 80$

Figure 5 : Small noise level ($\sigma = 0.5$). Large sparsity parameter ($d = \exp(-0.4)$).

Small noise, more friends

- Opinion convergence

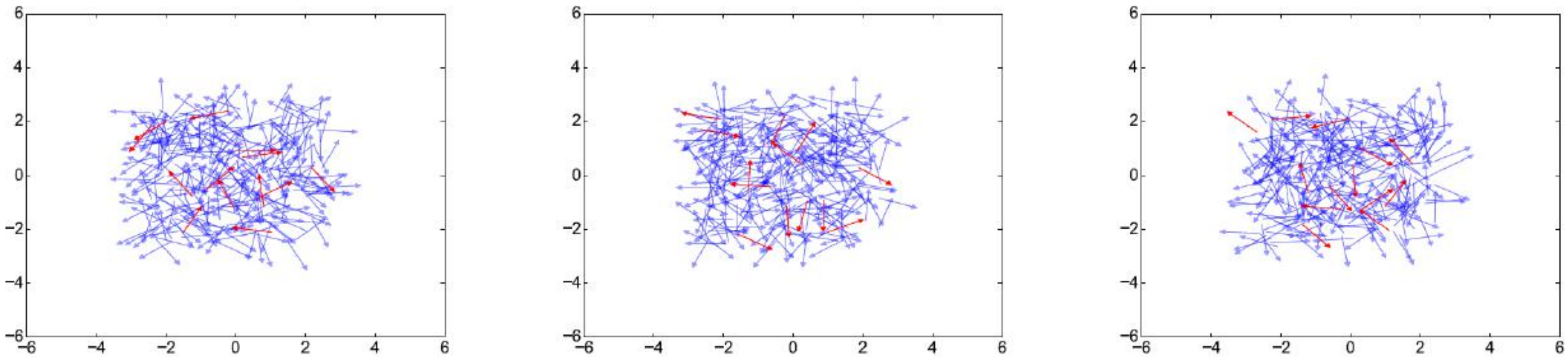


(a) $t = 0$ (init) (b) $t = 40$ (c) $t = 80$

Figure 6 : Small noise level ($\sigma = 0.5$). Small sparsity parameter ($d = \exp(-2.0)$).

Big noise

- Random

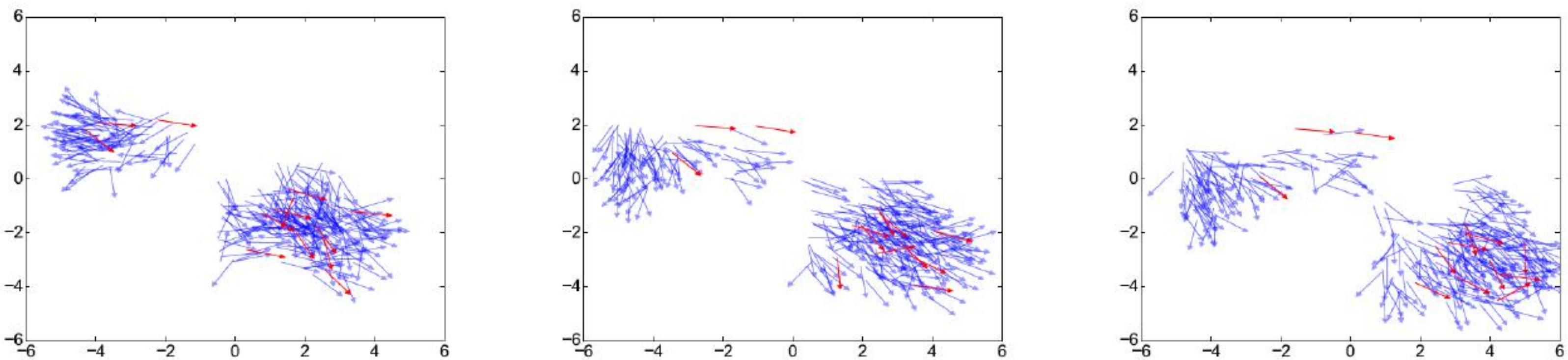


(a) $t = 0$ (init) (b) $t = 40$ (c) $t = 80$

Figure 7 : Large noise level ($\sigma = 2.0$). Large sparsity parameter ($d = \exp(-0.4)$).

Intervention

- How to alleviate opinion divergence?
  - Increase number of friends (diversity)
  - Strong opinion leaders



(a) $t = 80$ (b) $t = 120$ (c) $t = 160$

Figure 8 : Intervention: under the same setting: $\sigma = 0.5$, $d = \exp(-0.4)$).

Application

- Dataset: Co-sponsorship between legislators, extracted from congress voting record
  - https://www.govtrack.us

Notation	Value
Time period	1983-2016 (98*th*-114*th* congress meeting)
Number of time slices T	382 (month-based)
Number of legislators N	2,180
Number of co-sponsorship links	2.1 million

Application 1: Identify Opinion Leaders

- Learning popularity score: b

Rank	Name	Party-State	Time in Congress
1	Paul Simon	Democrat-IL	1975-1997
2	Jay Rockefeller	Republican-WV	1985-2015
3	John Kerry	Democrat-MA	1985-2013
4	Thomas Harkin	Democrat-IA	1975-2015
5	James Terry Sanford	Democrat-NC	1986-1993
6	Albert Gore	Democrat-TN	1983-1993
7	Kent Conrad	Democrat-ND	1987-2013
8	Edward Kennedy	Democrat-MA	1962-2009
9	Mitch McConnell	Republican-KY	1985-present
10	Frank Annunzio	Democrat-IL	1965-1993

Table 1: Popular legislators ranked by b in recent 34 years.

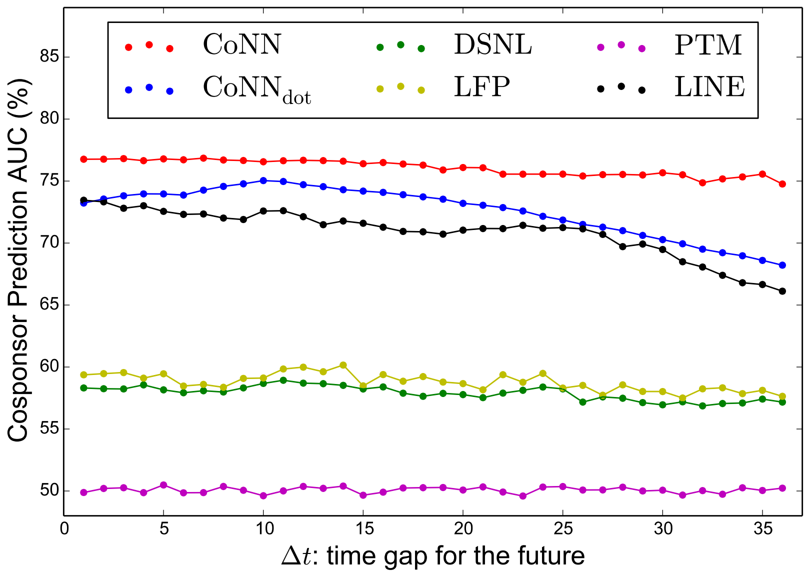
68th U.S. Secretary of State

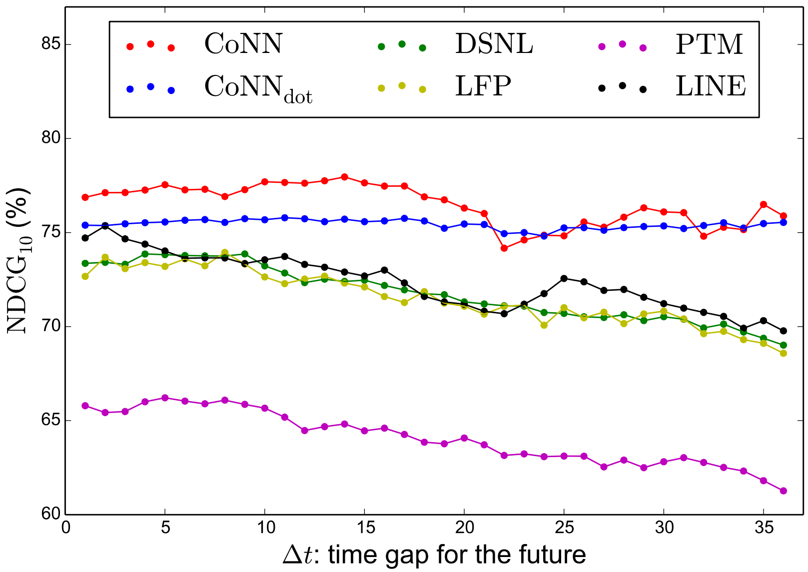
45th U.S. Vice President

the majority leader of
the Senate since 2015

Application 2: Predicting Co-Sponsorship

- Learning system-level parameters, and then run simulation





Outline

- Introduction
- Ideology Detection via Legislative Voting Network Mining
- Ideology Detection via Heterogeneous Types of Links
- User Stance Prediction via Joint Modeling of Text and Social Interactions
- The Co-Evolution Model for Social Network Evolving and Opinion Migration
- Summary

Summary

- Online behaviors are ubiquitous
  - Content
  - Interaction with other entities
  - Dynamic
- User stance prediction solution
  - Content rich information network mining
- Three case studies in this line
  - Issue-based ideology detection for congress people
  - Ideology detection for twitter users via heterogeneous network mining
  - The co-evolution model that can explain and predict the dynamics

Q & A

THANK YOU!