# SOCIAL MEDIA, GRAPHS AND COMMUNICATION

Paola Monachesi (Utrecht University)

# Where do I work?

Utrecht University, Trans 10

# Utrecht

# Uithof

# Where students live

# Where is Utrecht?

# What do I do?

- Language and social media
  - Dutch corpora and linguistic annotation
  - Extraction and formalization of information from SM to guide the learning process
  - Development of LT based functionalities to improve retrieval of learning material
  - Impact of physical space and online space on communication and information diffusion
  - Interdipendence between social structure and language

# Social Media

- Digital social dynamics match those in the physical world: friends are friends in both worlds

but differences:

- The number of people to interact with is not limited by distance or time

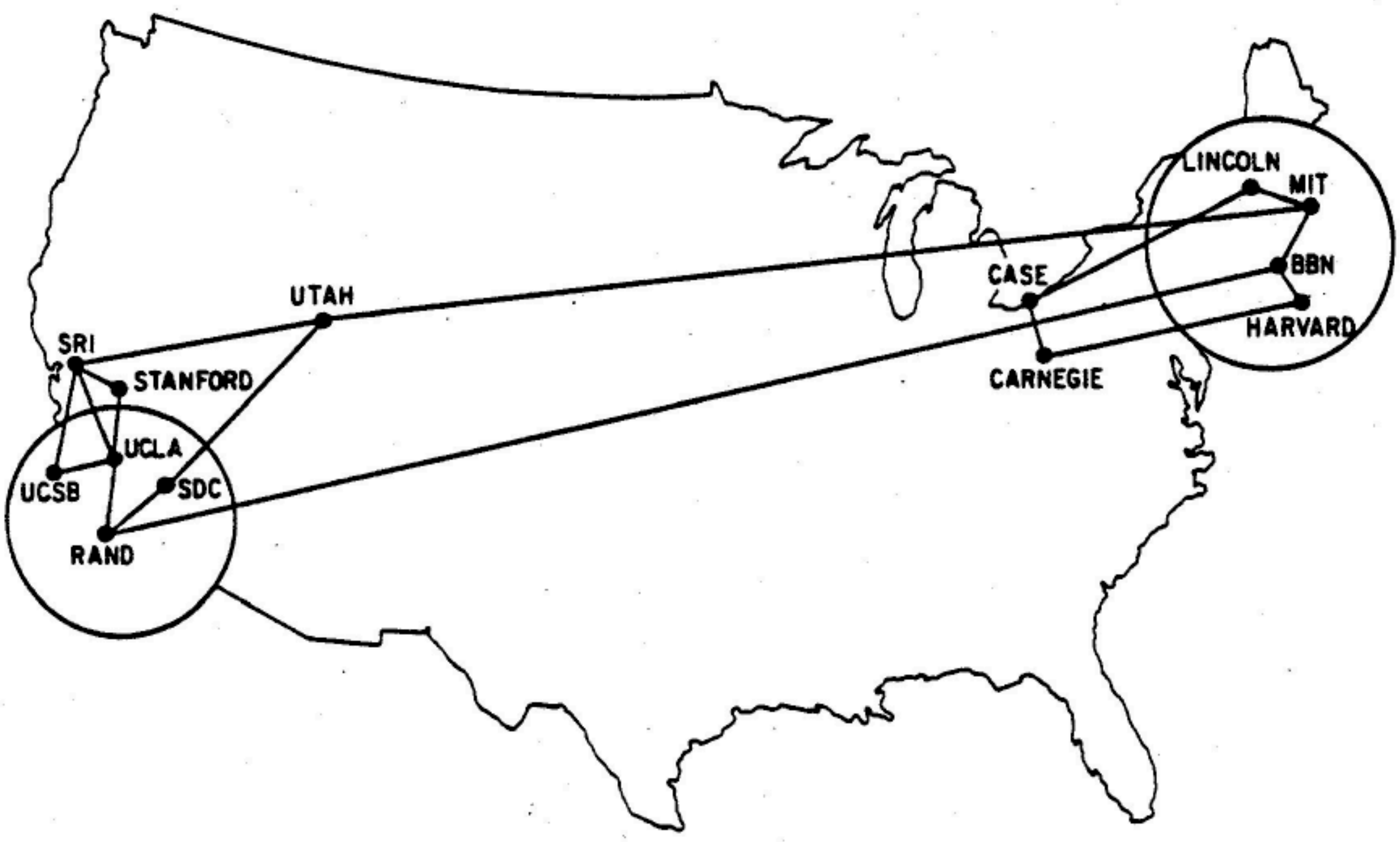
# Social Media

- Many new possibilities for us as users to communicate, interact, find and exchange information
- Many new possibilities for research
- Many new possibilities for enterprise

# Graphs

- Graph as mathematical models of network structure

# Communication Network



# Transportation network

# Social Networks

# Social Networks

- Paths
- Cycles
- Connectivity
- Connected components
- Length of path
- Distance
- Small world phenomenon – six degree of separation

# Twitter analysis

- SM influence politics and trigger political communication
- Tendency to polarization and segregation
- Risk of not being exposed to diversity: online communities
- Opinions can become more extreme
- Conover et al. (2011) “ *Political polarization on Twitter*”. AAAI conference on Weblogs and SM.
- Analysis of Twitter data for political discourse

# Communities and communication

- Examine networks of political communication
- How do we do it?
- Which tools do we use?
- Which results do we expect?
- Relevance of:
  - Retweets
  - Mention
  - Hashtags (#)

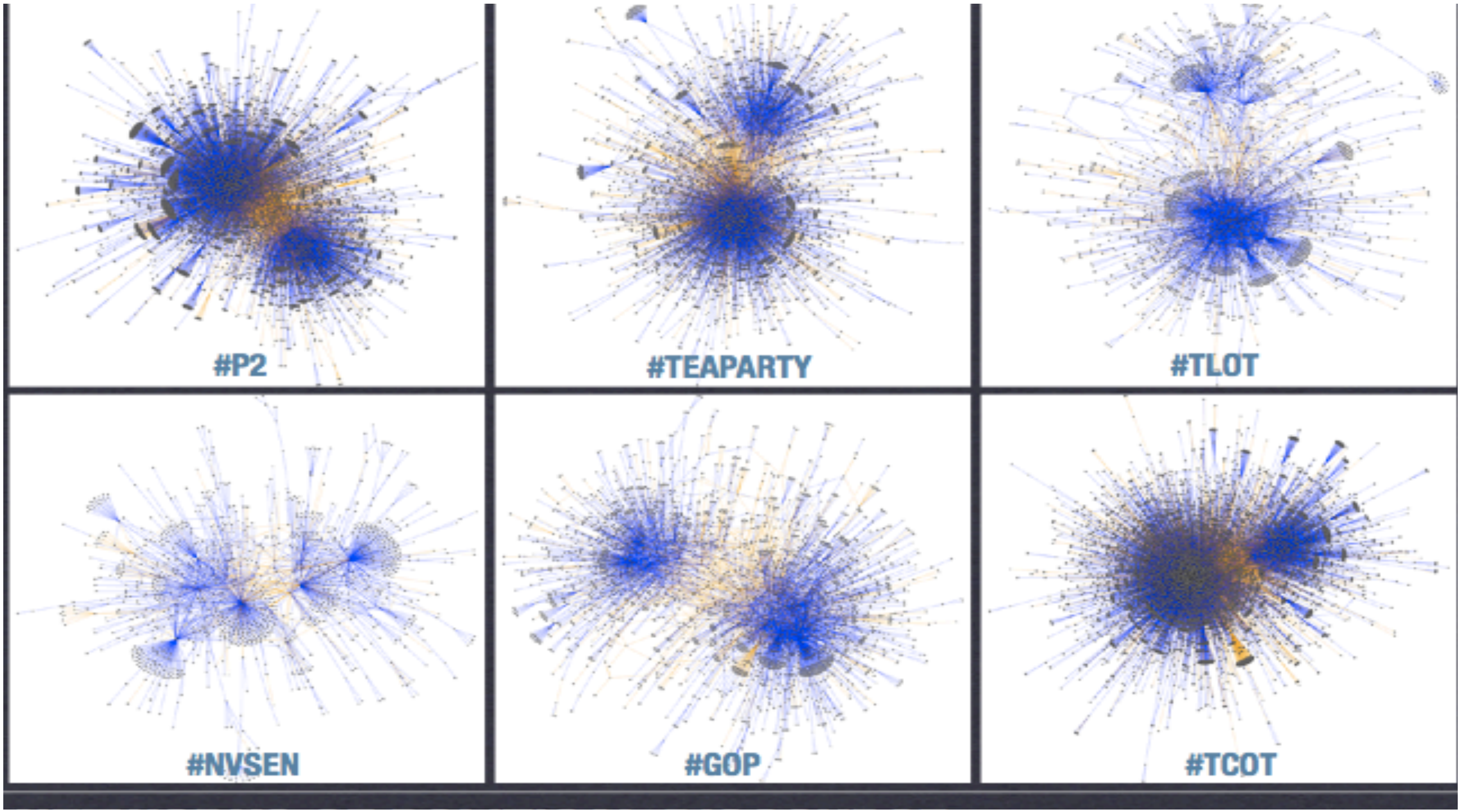
# Methodology

- Creation of a network and text data set from Twitter
- Cluster analysis of network and properties of retweet and mentions
- Manual classification of Twitter users to understand the nature of the networks (i.e analysis of users)
- Interpretation of the community structures

# Methodology

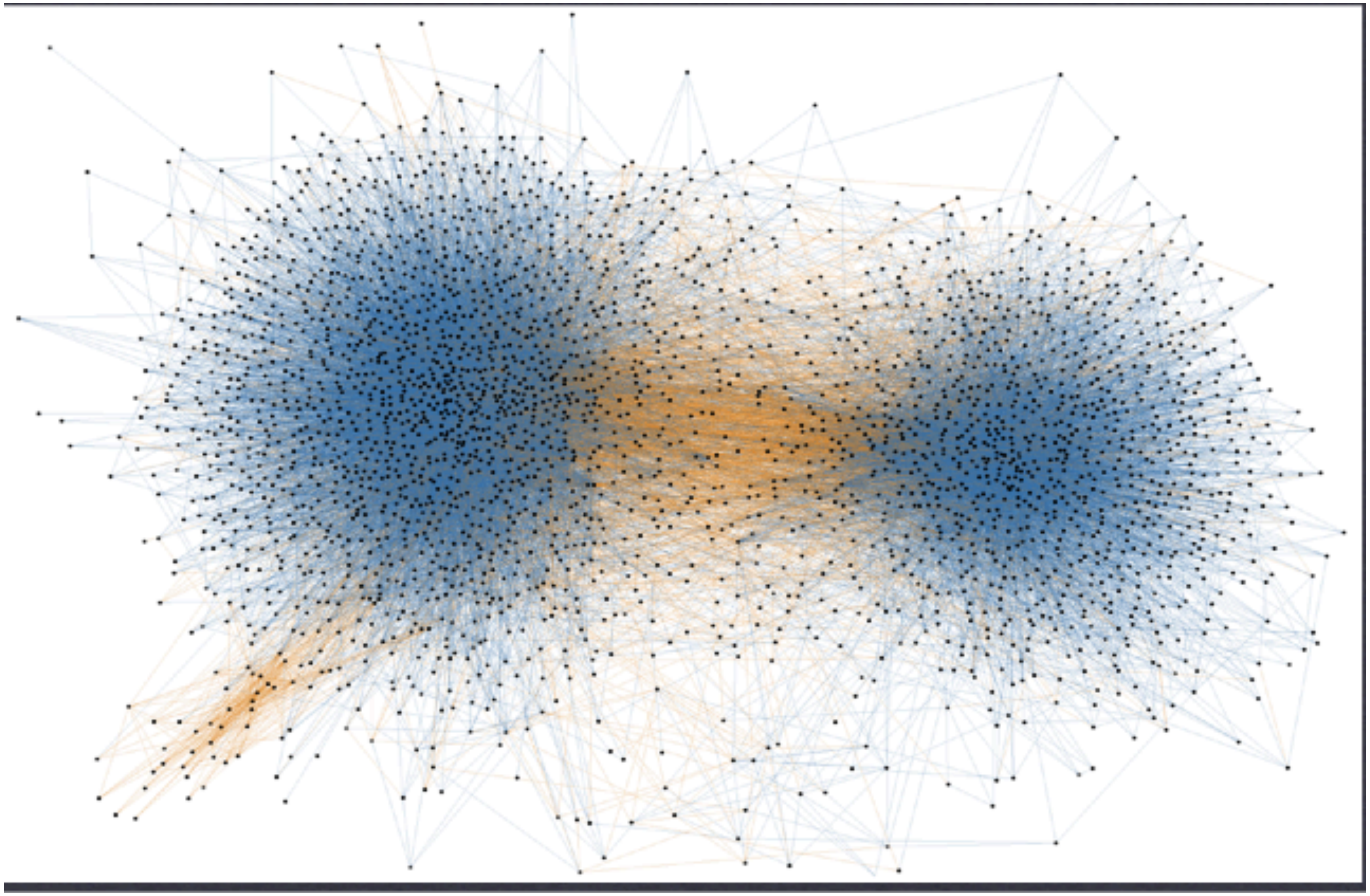
- Network analysis is not enough
- Use of qualitative analysis from social sciences
- Manual annotation of users political trends to get insights into the data

# Networks of political discourse
(atomic structure)

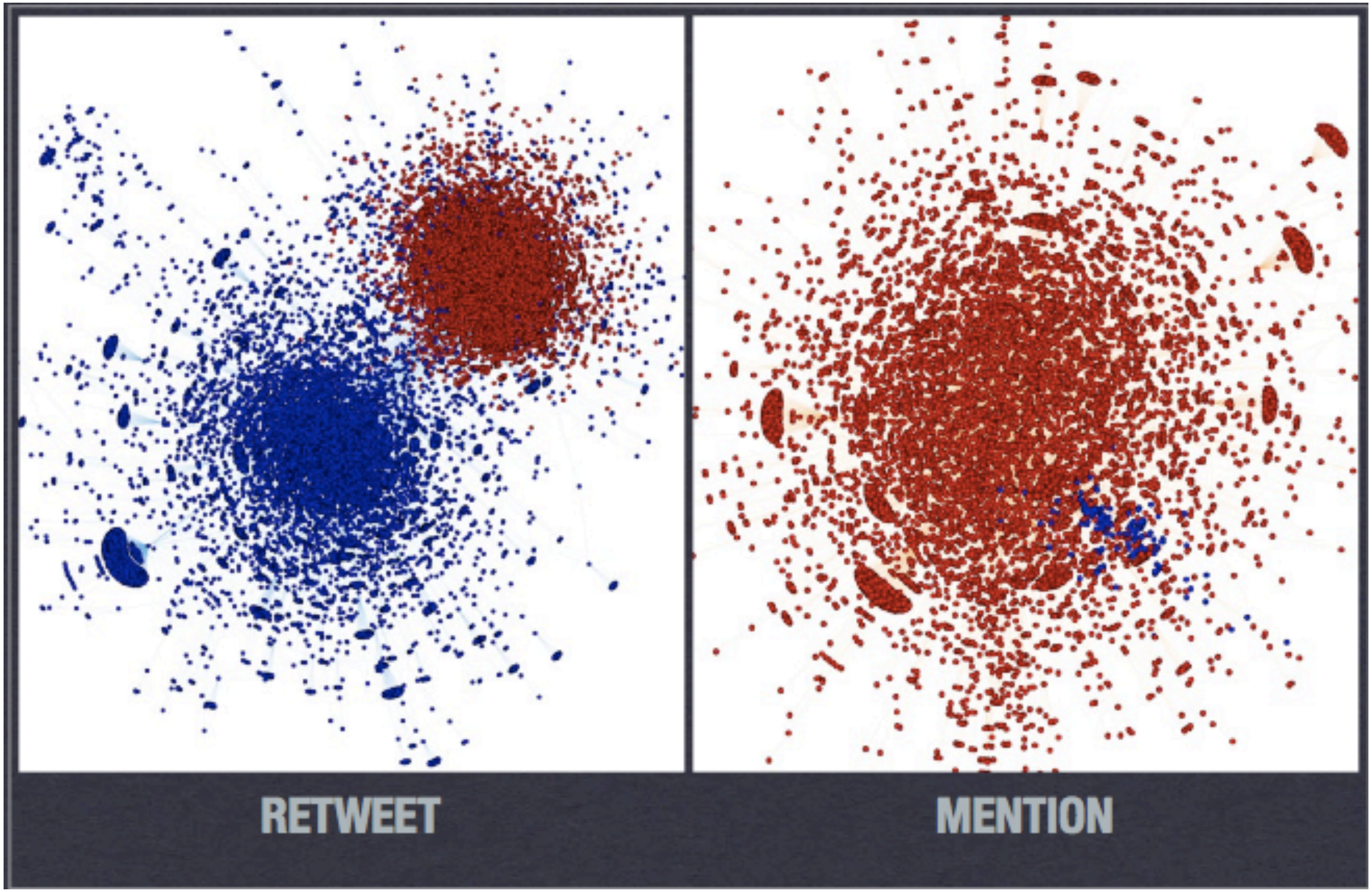


# Network of political discourse

## Aggregate structure



# Multi-mode communication



# Observations

- Impact of SM on political communication
- Retweets: segregation
- Mention: interaction among different opinions triggered by political motivated individuals through #
- Use of #: expose users to content they would not choose in advance

# Findings

- Different use of rewteet and mention in Twitter political communication and in the way information flows;
- Not accidental but the result of political people that inject content through an appropriate use of hashtags;
- Ideologically opposed users are the target, they are not going to rebroadcast the tweet but use of mentions

# Twitter

- Twitter: microbloging site
- 140 charcters = tweets
- Interaction:
  - Retweets: rebroadcast content of other users
  - Mentions: address a user through the public feed (i.e. any Twitter update that contains "@username" anywhere in the body of the Tweet)
  - Hashtags: metadata about a topic or intended audience

# Data used

- Analysis based on data collected through the Twitter api during 6 weeks before US congress midterm elections in 2010
- 355 million tweets
- Need to make a selection
- How?

# Identify political content

- Find tweets that contain at least one political #
- Tag co-occurence discovery
- Use of seed tags (i.e. #p2, #tcot)
- Identify set of # that co-occur in at least one tweet
- Results ranked using Jaccard Coefficient:

$$\sigma(S,T) = \frac{|S \cap T|}{|S \cup T|}. \qquad (1)$$

- Threshold of 0.005

# Resulting data

- Identify 66 unique #
- (11 excluded – ambiguos)
- Total 252300 tweets

# Identify political content

Table 1: Hashtags related to #p2, #tcot, or both. Tweets containing any of these were included in our sample.

| | |
|---|---|
| Just #p2 | #casen #dadt #dc10210 #democrats #du1 #fem2 #gotv #kysen #lgf #ofa #onenation #p2b #pledge #rebelleft #truthout #vote #vote2010 #whyimvotingdemocrat #youcut |
| Both | #cspj #dem #dems #desen #gop #hcr #nvsen #obama #ocra #p2 #p21 #phnm #politics #sgp #tcot #teaparty #tlot #topprog #tpp #twisters #votedem |
| Just #tcot | #912 #ampat #ftrs #glennbeck #hhrs #iamthemob #ma04 #mapoli #palin #palin12 #spwbt #tsot #tweetcongress #ucot #wethepeople |

# Identify political content

Table 2: Hashtags excluded from the analysis due to ambiguous or overly broad meaning.

| | |
|---|---|
| Excl. from #p2 | #economy #gay #glbt #us #wc #lgbt |
| Excl. from both | #israel #rs |
| Excl. from #tcot | #news #qsn #politicalhumor |

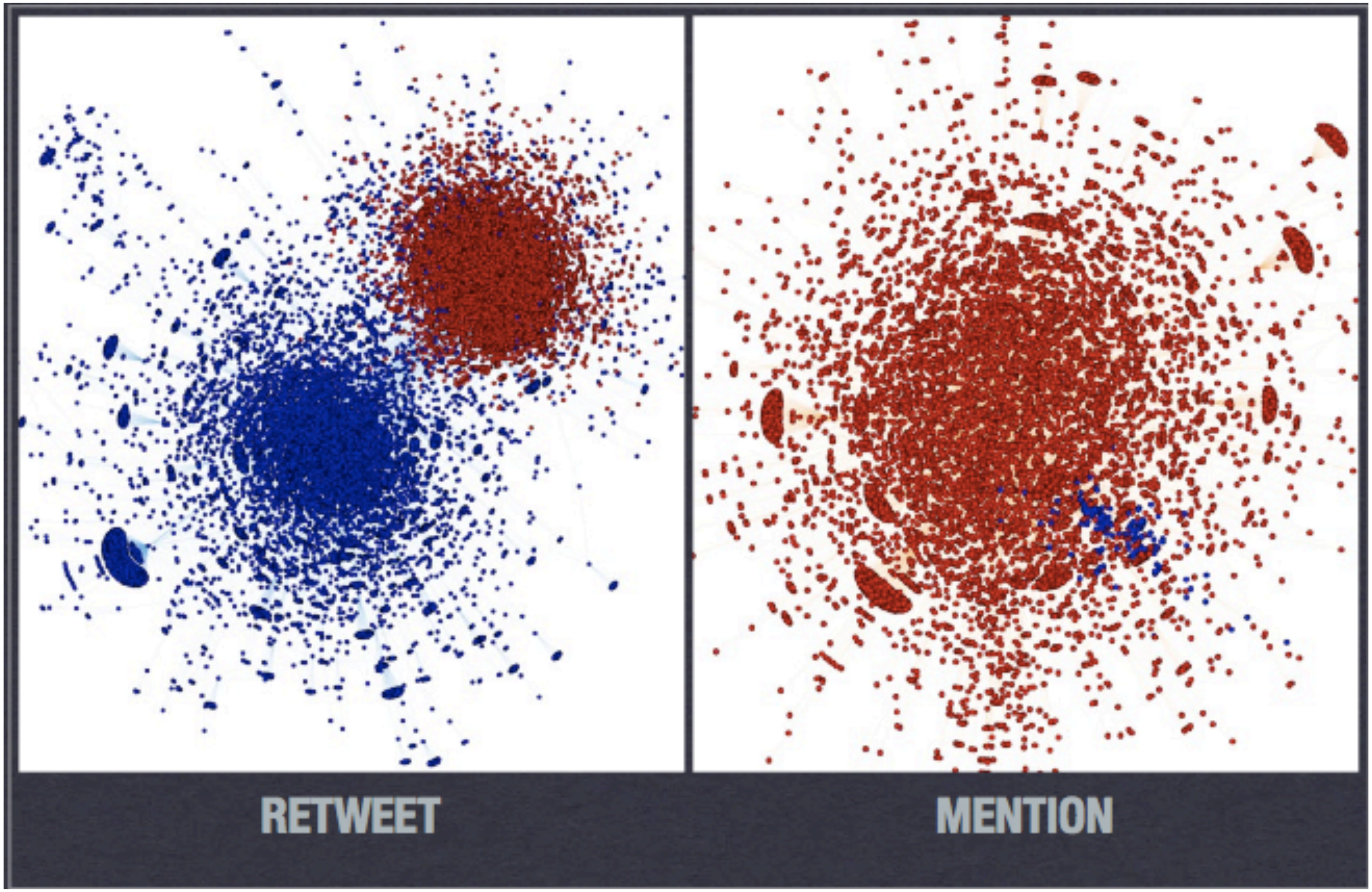
# Political communication Networks

- Construct a network based on the retweets and mention
- Information flowing from A to B
- RN: total ~45k nodes, ~23k non isolated nodes, largest connected component ~18k nodes
- MN: total ~17k nodes, ~10k non isolated nodes, largest connected component ~7k nodes

# Community structure

- Community detection: label propagation method
  - Assign arbitrary cluster membership to each node
  - Iteratively update each node's label on the basis of the label that is shared by most of its neighbors
- RN: 2 clusters of users that propagate content within their community
- MN: we don't find these clusters

# Multi-mode communication

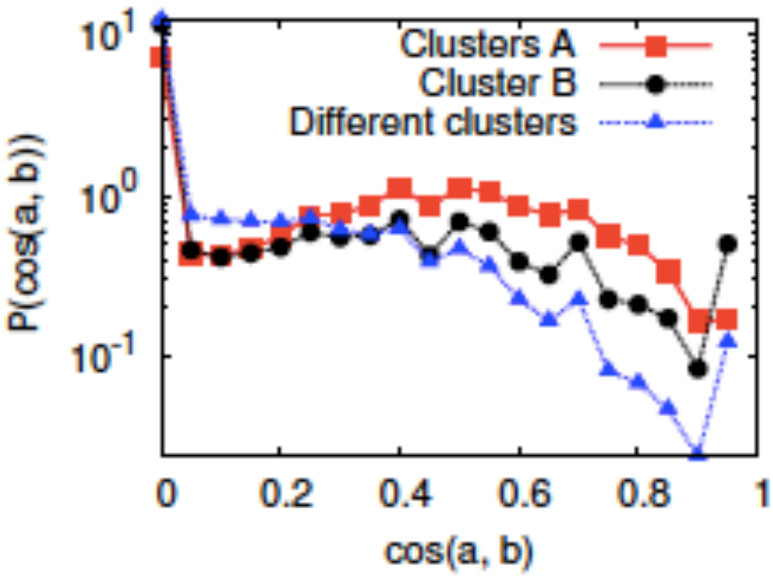


# Content analysis

- Clustering based on network properties
- Are these clusters related to the *content* of the discussions involved?
  - Associate users with a profile vector containing # in own tweets weighted by frequency
  - Compute cosine similarity between pair of users profiles within the same cluster and in different clusters
- RN: users in cluster A have more similar profiles than users in cluster B
- MN: this is not the case

# Cosine similarities among user profiles

| | Retweet | Mention |
|---|---|---|
| A↔A | 0.31 | 0.31 |
| B↔B | 0.20 | 0.22 |
| A↔B | 0.13 | 0.26 |



# Political polarization

- Do clusters in the retweet network correspond to users with similar political views?
- Qualitative content analysis
- Identify whether the tweet of a given user expresses a left, right or undecidable identity
- Author annotates 1000 random users
- Non author annotates 200 from the set of 1000 users
- Check agreement between annotation

# Annotation Agreement

- Kappa coefficient

$$\kappa = \frac{P(\alpha) - P(\epsilon)}{1 - P(\epsilon)}$$

- P(a)= observed rate of agreement
- P(e) = expected rate of random agreement given the relative frequency of each class label
- K=0.80 (left wing)
- K= 0.82 (right wing)
- K= 0.42 (undecidable)

# Political divisions

Table 4: Partisan composition and size of network clusters as determined by manual inspection of 1,000 random user profiles.

| Network | Clust. | Left | Right | Undec. | Nodes |
|---|---|---|---|---|---|
| Retweet | A | 1.19% | 93.4% | 5.36% | 7,115 |
| | B | 80.1% | 8.71% | 11.1% | 11,355 |
| Mention | A | 39.5% | 52.2% | 8.18% | 7,021 |
| | B | 9.52% | 85.7% | 4.76% | 154 |

# Cross ideological interaction

- Users are likely to interact with other with whom they agree (retweet)
- More cross ideological interaction in the mention network

# Content injection

- Use of # that target different politically opposed audiences
- Expose users to different information
- No retweet, but use of mention to reply

# Use of tags by communities

| Rank | Hashtag | Left | Right | Valence |
|---|---|---|---|---|
| 1 | #tcot | 2,949 | 13,574 | 0.384 |
| 2 | #p2 | 6,269 | 3,153 | -0.605 |
| 3 | #teaparty | 1,261 | 5,368 | 0.350 |
| 4 | #tlot | 725 | 2,156 | 0.184 |
| 5 | #gop | 736 | 1,951 | 0.128 |
| 6 | #sgp | 226 | 2,563 | 0.694 |
| 7 | #ocra | 434 | 1,649 | 0.323 |
| 8 | #dems | 953 | 194 | -0.818 |
| 9 | #twisters | 41 | 990 | 0.843 |
| 10 | #palin | 200 | 838 | 0.343 |
| | Total | 26,341 | 53,880 | |

# Example

User A: Please follow @Username for an outstanding progressive voice! #p2 #dems #prog #democrats #tcot

User B: Couple Aborts Twin Boys For Being Wrong Gender..http://bit.ly/xyz #tcot #hhrs #christian #tlot #teaparty #sgp #p2 #prolife

# Political valence

- It encodes the relative prominence of a tag among left and right wing users

$$V(t) = 2\,\frac{N(t,R)/N(R)}{[N(t,L)/N(L)] + [N(t,R)/N(R)]} - 1 \quad (4)$$

- N(t,R) = number of occurences of tag (t) produced by right wing users
- N(t,L) = same for left wing users
- N(R) = total number of occurrences of all tags in tweets by right wing users
- N(L) same for left wing users
- Constants used to bound the measure between -1 for tag used by the left and +1 for tag used by the right

# Relevance of paper

- Analysis
  - Verification of hypothesis possible
  - Identification of different uses of communication means (retweet, mention, #)
  - Information sharing: within the same community (retweet)
  - Inetgrate network analysis with content analysis

# Relevance of paper

- Methodology
  - Data extraction (co-occurance of #)
  - Network construction
  - Clustering analysis: community detection
  - Content analysis: # analysis to identify similarity of users within cluster
  - Qualitative content analysis: annotation + classification to identify left and rightwing users

# Other uses

- Can we use this methodology to discover communication behavior in other communities?
- Which ones?

# Relevant resources

- The paper:

http://truthy.indiana.edu/site_media/pdfs/conover_icwsm2011_polarization.pdf

- Talk by the author: http://videolectures.net/icwsm2011_conover_polarization/
- The system used and data:

http://truthy.indiana.edu/