Social Media & Text Analysis

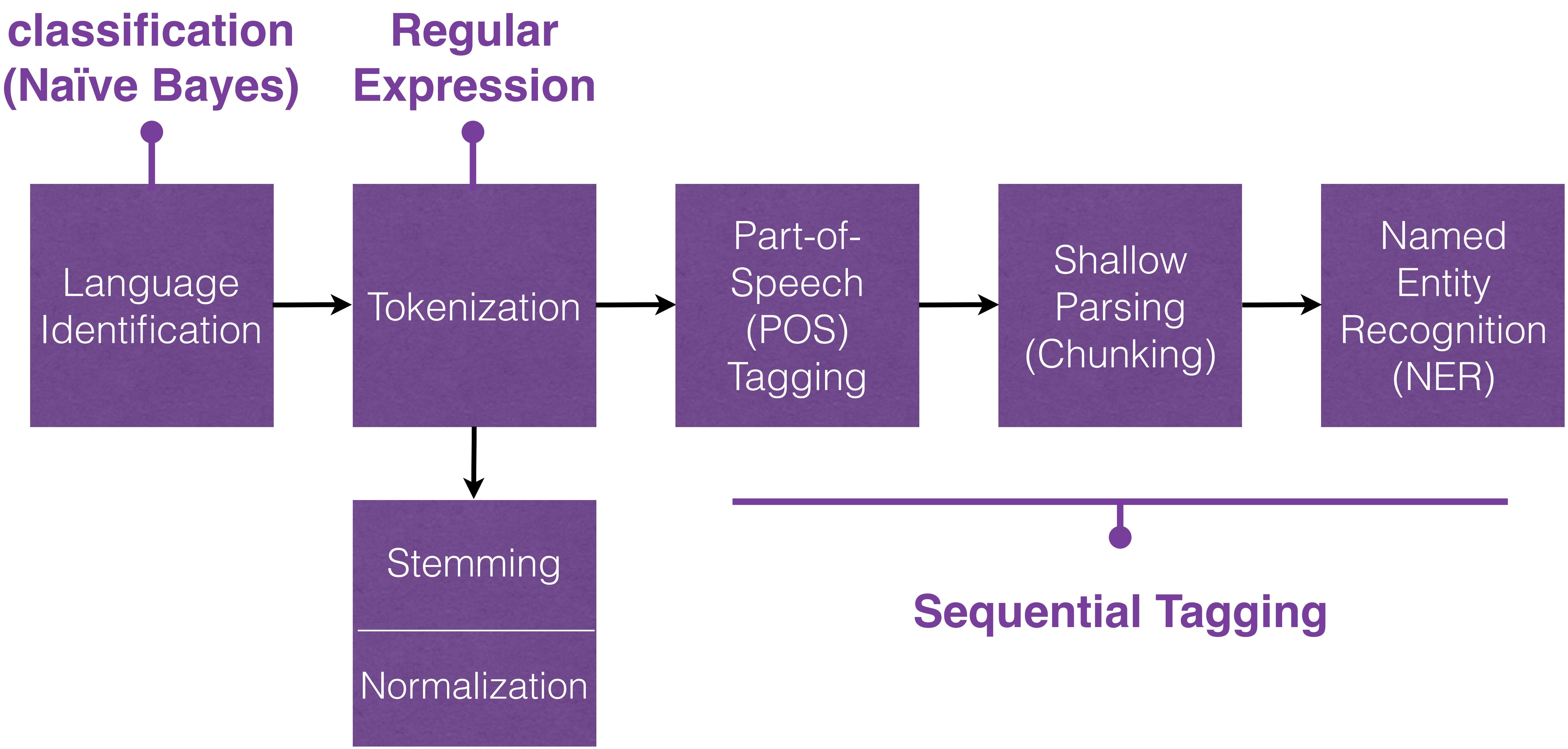
lecture 6 - Automatic Summarization for Twitter

Follow @cocoweixu

Instructor: Wei Xu

Website: socialmedia-class.org

[Recap] NLP Pipeline



Timeline of NLP on Microblogs

2010

- Dialogue Modeling (Ritter et al.)
- First Story Detection (Petrovic et al.)
- Geographic Variation (Eisenstein et al.)

2011

- **Named Entity Recog. (Ritter et al.)**
- **POS Tagging (Gimpel et al.)**
- **Normalization (Han and Baldwin)**
- Summarization (Liu et. al)

2012

- **Open-Domain Event Extraction (Ritter et al.)**
- Censorship Detection (Bamman et al.)

Timeline of NLP on Microblogs

2013

- **Summarization (Xu et al.)**
- **Normalization (Xu et al.)**
- Machine Translation (Ling et. al.)
- POS (Owoputi et al.)

2014

- **Paraphrase Extraction (Xu et. al.)**
- Parsing Weibo (Wang et. al.)
- Parsing Twitter (Kong et. al.)

2015

- **POS (Cherry and Guo)**
- Dialogue Modeling (Sordoni et al.)

Twitter Usage / Company Facts

302M
Monthly active users

500M
Tweets sent per day

80%
Active users on mobile

3,900
Employees around the world

35+
Offices around the world

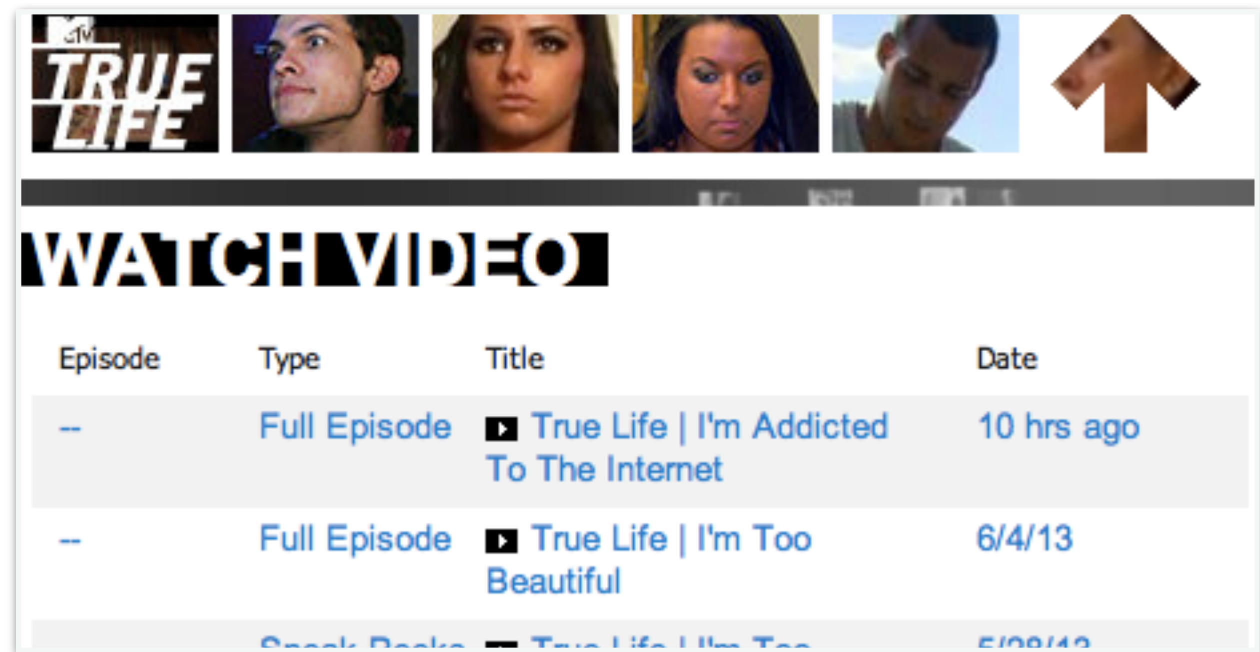
77%
Accounts outside the U.S.

33
Languages supported

50%
Employees are engineers

All numbers approximate as of March 31, 2015.

Summarization



SUMMARY:

I'm watching **true life** **"I'm addicted to Internet"** … while I'm on mine lol

Okay these girls on **True Life** **I'm Too Beautiful** are not that pretty

Summarization

- Given a (or a set) of documents, generate a short summary

- Given a (large) set of topically and temporally clustered tweets, select a few representative tweets as the summary.

Previous Work

Selected Work	Size of Input	Length of Summary
Wei et al. (2012)	average 10k tweets	10 tweets
Inouye & Kalita (2011)	approximately 1500 tweets	4 tweets ❖
Rosa et al. (2011)	average 410 tweets	1, 5, 10 tweets
Liu et al. (2011)	average 1.7k tweets	about 2 or 3 tweets ★
Takamura et al. (2011)	2.8k - 5.2k tweets	26 - 41 tweets ★

❖ Human annotators strongly prefer different numbers of tweets in a summary for different topics.

★ Used the length of human reference summaries to decide the length of system outputs, which information is not available in practice.

Wei Xu, Alan Ritter, Ralph Grishman.
“A Preliminary Study of Tweet Summarization using Information Extraction” in LASM (2014)

SumBasic

- Intuition:

  words occurring frequently in the documents occur with higher probability in the human summaries than words occurring less frequently

SumBasic

- a very simple but strong summarization algorithm [Nenkova and Vanderwende, 2005]
- Intuition:

  words occurring frequently in the documents occur with higher probability in the human summaries than words occurring less frequently

SumBasic

- Step 1: computes the probability of each word ***w*** :

$$P(w) = \frac{n(w)}{\sum_i w_i}$$

- Step 2: computes the salience score of each sentence ***S*** :

$$Score(S) = \sum_{w \in S} \frac{P(w)}{|\{w \mid w \in S\}|}$$

- Step 3: pick the highest scored sentence into summary

- Step 4: for each word in sentences chosen at step 3, update their probability:

$$P_{new}(w) = P_{old}(w) \cdot P_{old}(w)$$

- Step 5: repeat Step 2~4 until reach desired length of summary

Varied-length Summary

- For a set of topically clustered tweets, amount of information varies greatly:
  - from very repetitive to very discrete
  - e.g.

album release of a less notable singer
vs.
album release of a famous/controversy singer

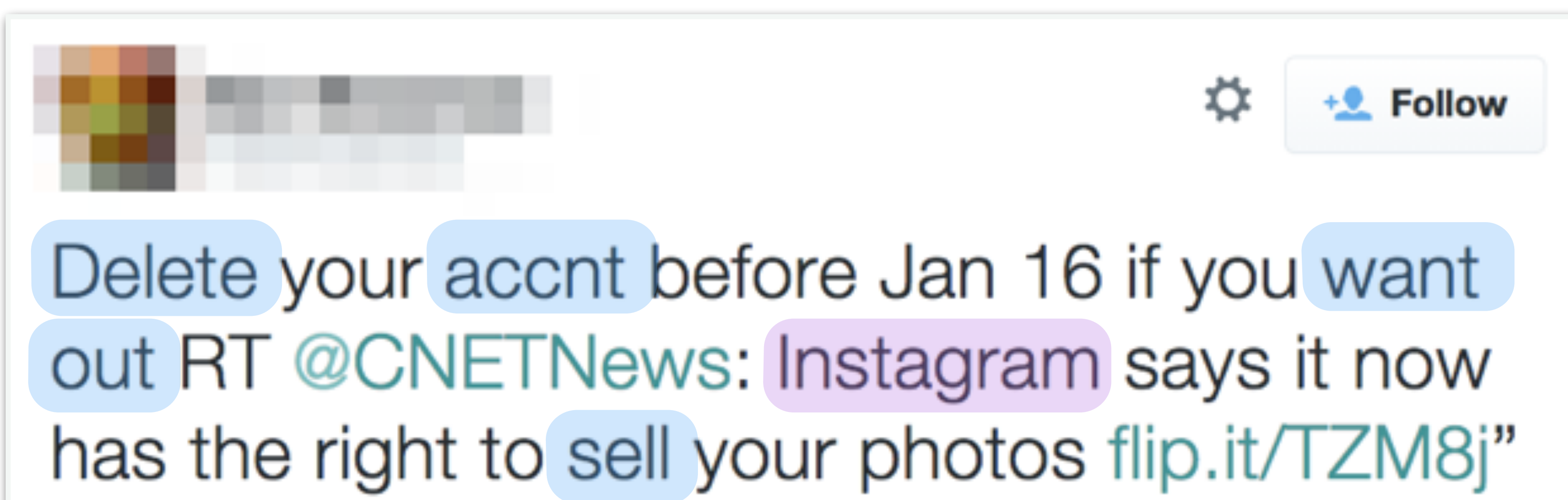
Wei Xu, Alan Ritter, Ralph Grishman.
"A Preliminary Study of Tweet Summarization using Information Extraction" in LASM (2014)

Information Extraction (IE)

- Named Entity [Ritter et al. 2011]
- Event Phrases [Ritter et al. 2012]

Information Extraction (IE)

- Named Entity [Ritter et al. 2011]
- Event Phrases [Ritter et al. 2012]

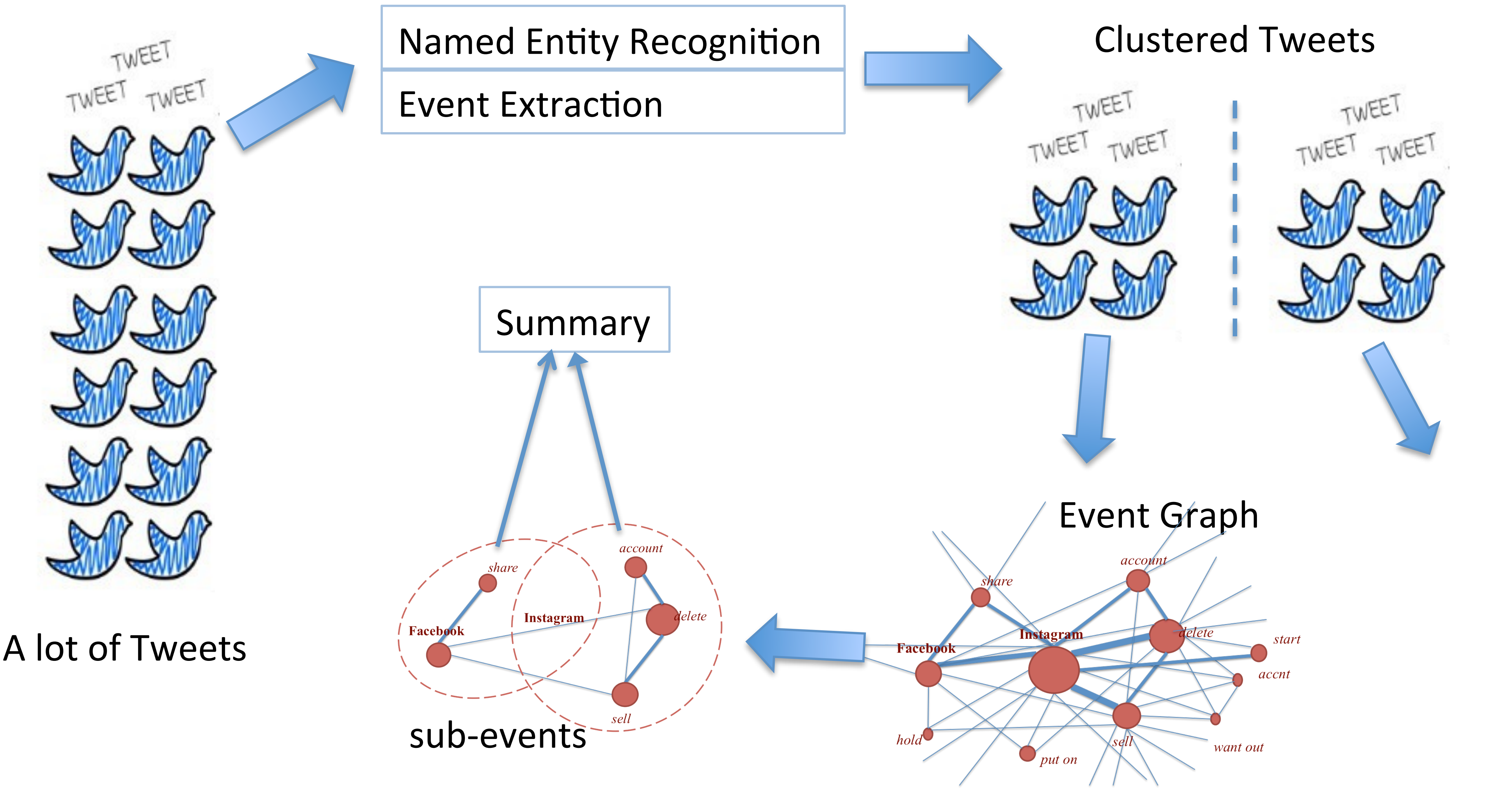


Research Questions

- What is the perfect length of multi-tweet summary?
- Will IE help summarization on Twitter?
  - noisy text: performance of IE?
  - short context: still need in-depth event analysis?
  - redundant: is word enough?

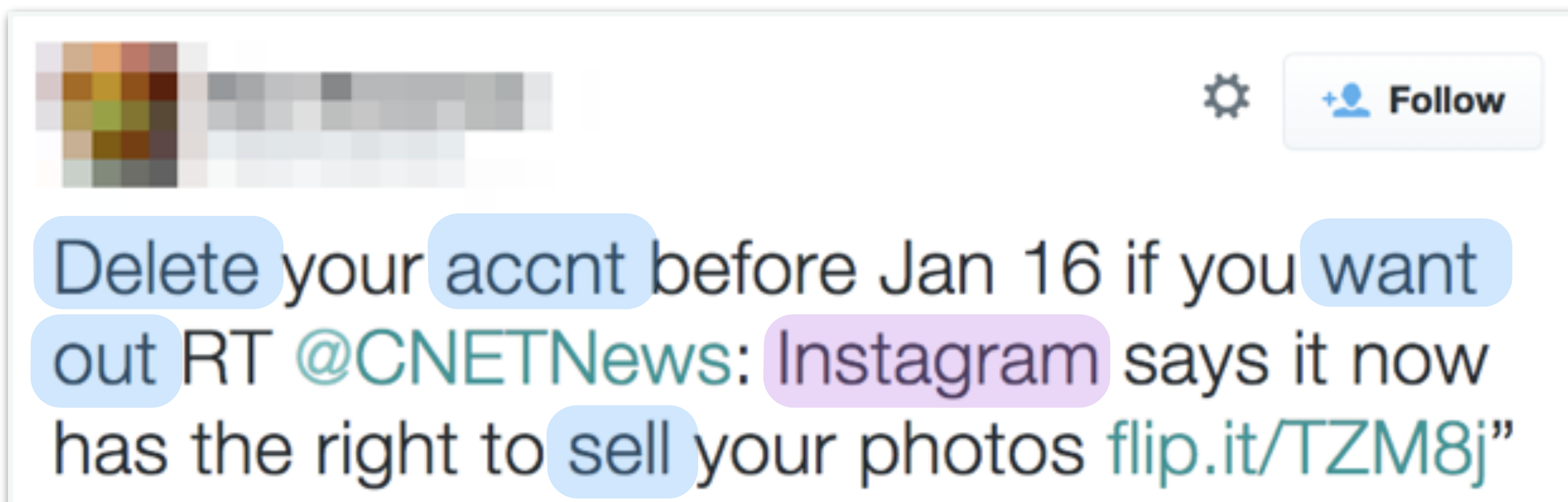
Wei Xu, Alan Ritter, Ralph Grishman.
"A Preliminary Study of Tweet Summarization using Information Extraction" in LASM (2014)

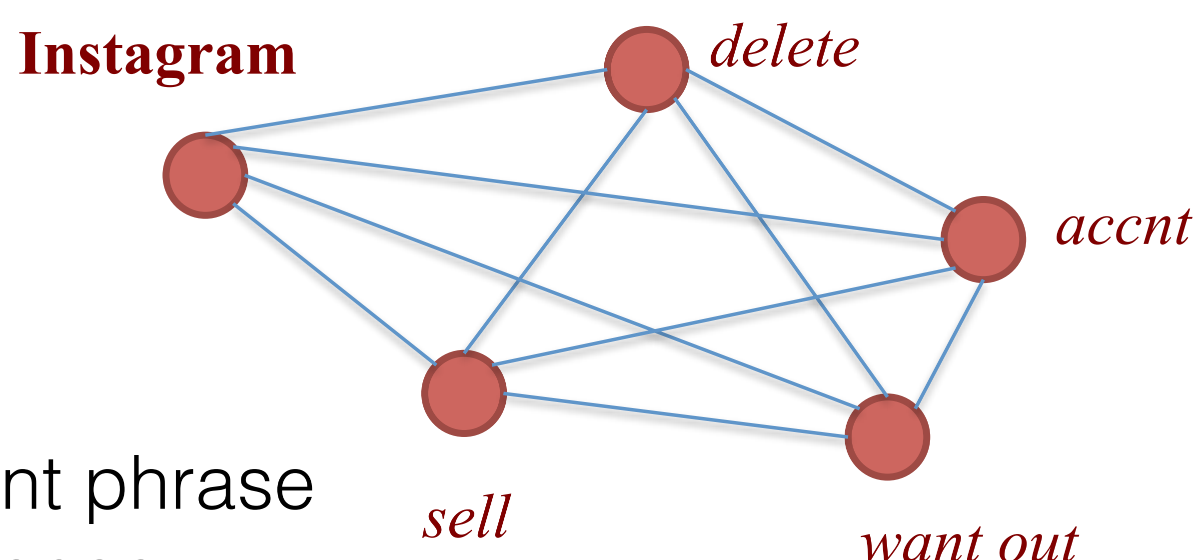
System Overflow



Wei Xu, Alan Ritter, Ralph Grishman.
"A Preliminary Study of Tweet Summarization using Information Extraction" in LASM (2014)

Event Graph

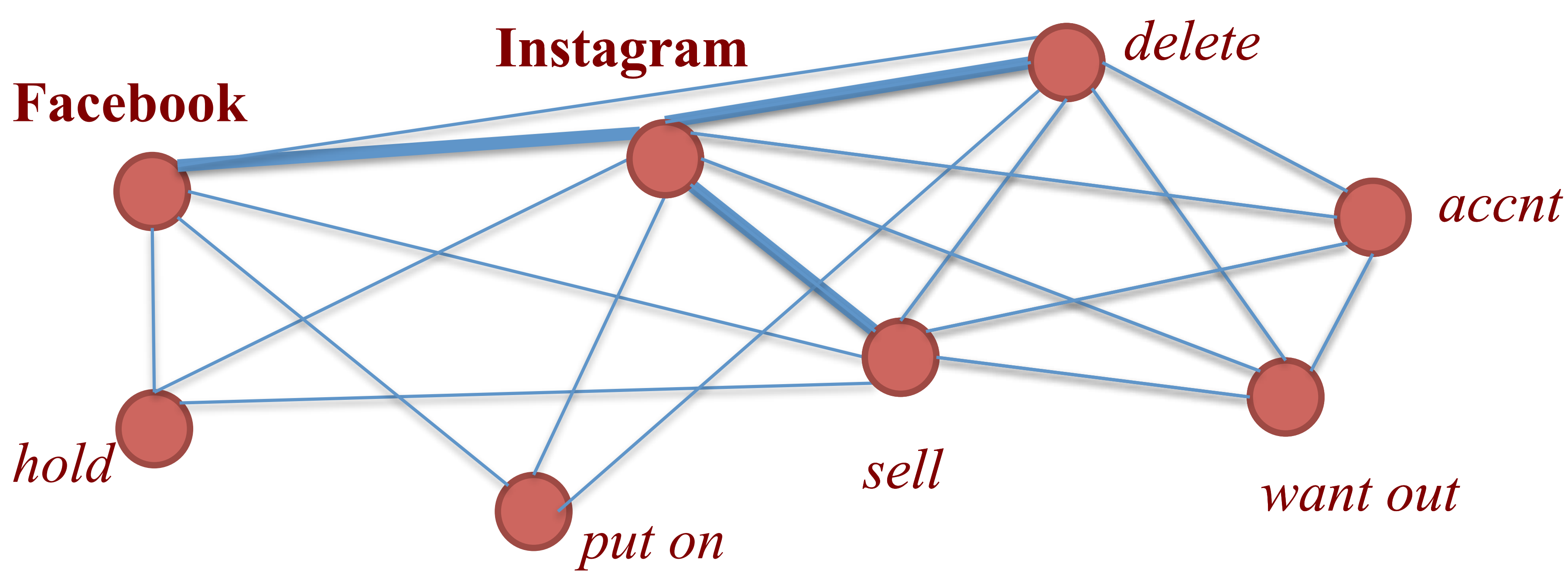




Node - named entities + event phrase
Edge (weighted) - co-occurrence

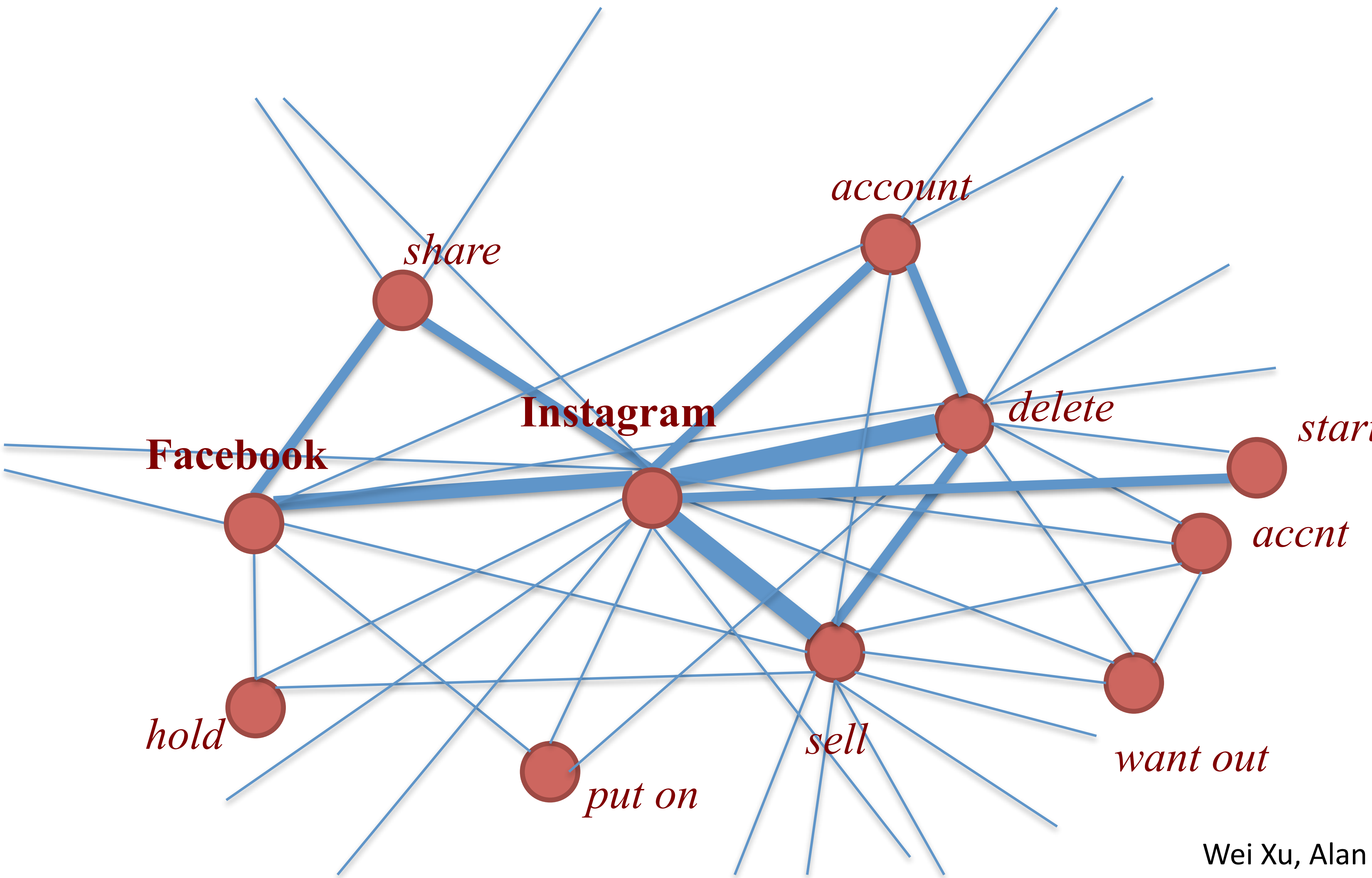
Wei Xu, Alan Ritter, Ralph Grishman.
"A Preliminary Study of Tweet Summarization using Information Extraction" in LASM (2014)

Event Graph



Wei Xu, Alan Ritter, Ralph Grishman.
"A Preliminary Study of Tweet Summarization using Information Extraction" in LASM (2014)

Event Graph



Wei Xu, Alan Ritter, Ralph Grishman.
“A Preliminary Study of Tweet Summarization using Information Extraction” in LASM (2014)

PageRank

- a graph-based ranking algorithm
- a trademark of Google
- Idea: web surfing / random walk

  The importance of a webpage is defined recursively and depends on the number and importance of all webpages that link to it.

- also used for local graph partitioning

PageRank

- Salience score of nodes:

$$Score(u) = (1-d) + d \times \sum_{v \in Adj(u)} \frac{Score(v)}{|Adj(v)|}$$

adjacent nodes

- directed graph
- iterate towards converge
- initial rank of node does not matter
- only edges matter
- total weight of the graph stays the same

PageRank → Event Rank

- Salience score of nodes:

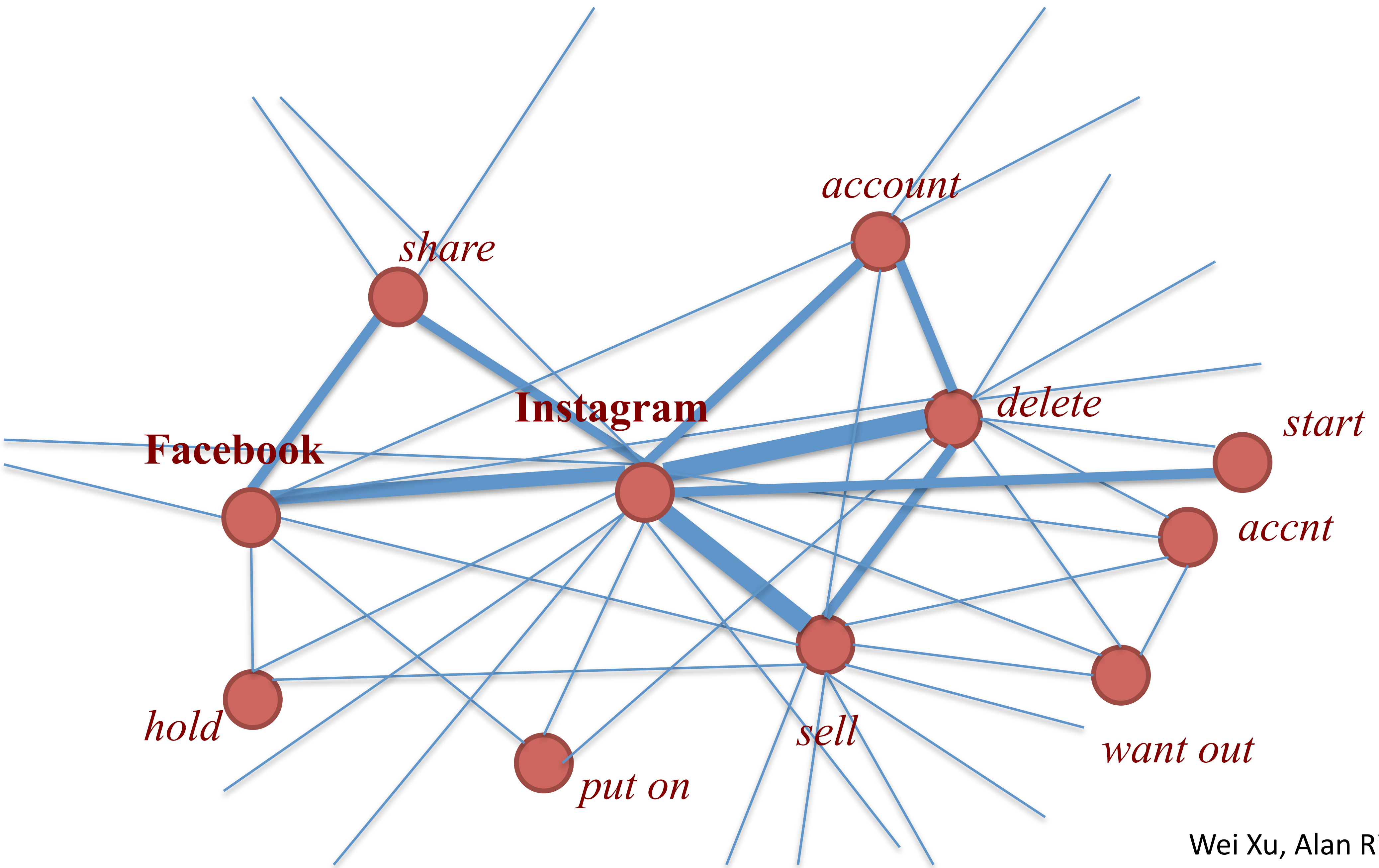
$$Score(u) = (1-d) + d \times \sum_{v \in Adj(u)} \frac{e_{uv} \times Score(v)}{\sum_{w \in Adj(v)} e_{vw}}$$

adjacent nodes

  - undirected graph
  - iterate towards converge
  - initial rank of node does not matter
  - only edges and their weights matter
  - total weight of the graph stays the same

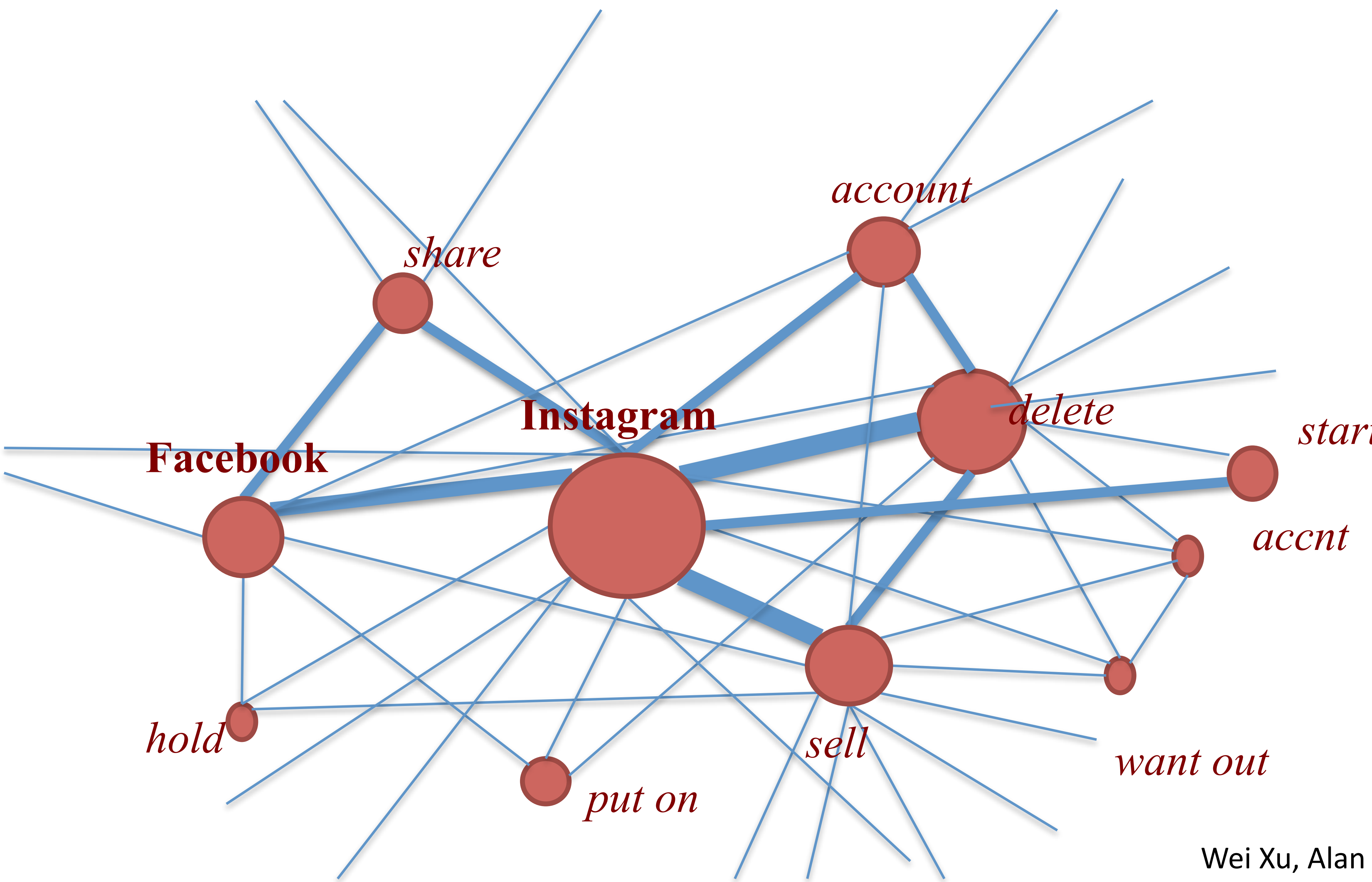
Wei Xu, Alan Ritter, Ralph Grishman.
"A Preliminary Study of Tweet Summarization using Information Extraction" in LASM (2014)

Graph Ranking



Wei Xu, Alan Ritter, Ralph Grishman.
"A Preliminary Study of Tweet Summarization using Information Extraction" in LASM (2014)

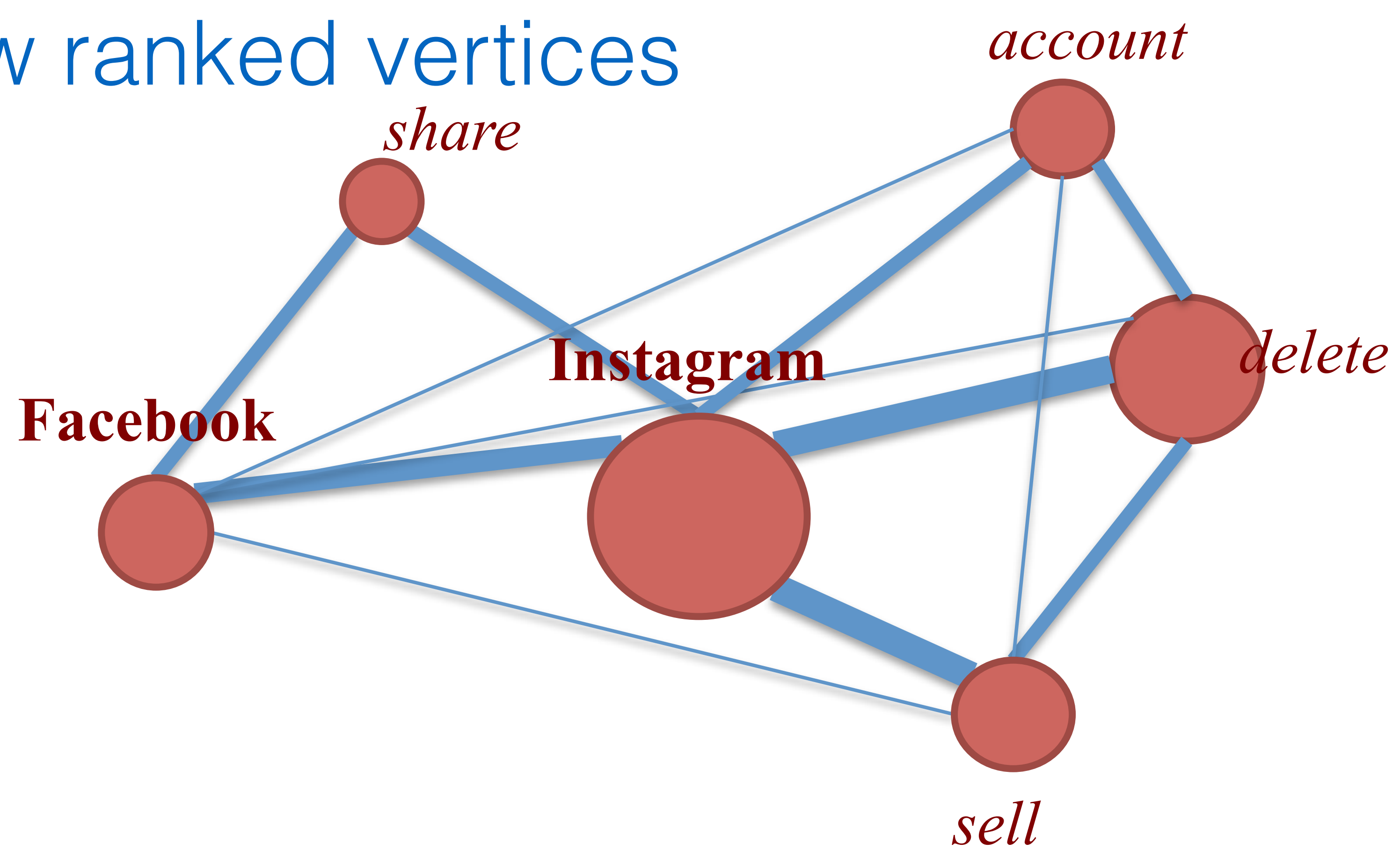
Graph Ranking



Wei Xu, Alan Ritter, Ralph Grishman.
"A Preliminary Study of Tweet Summarization using Information Extraction" in LASM (2014)

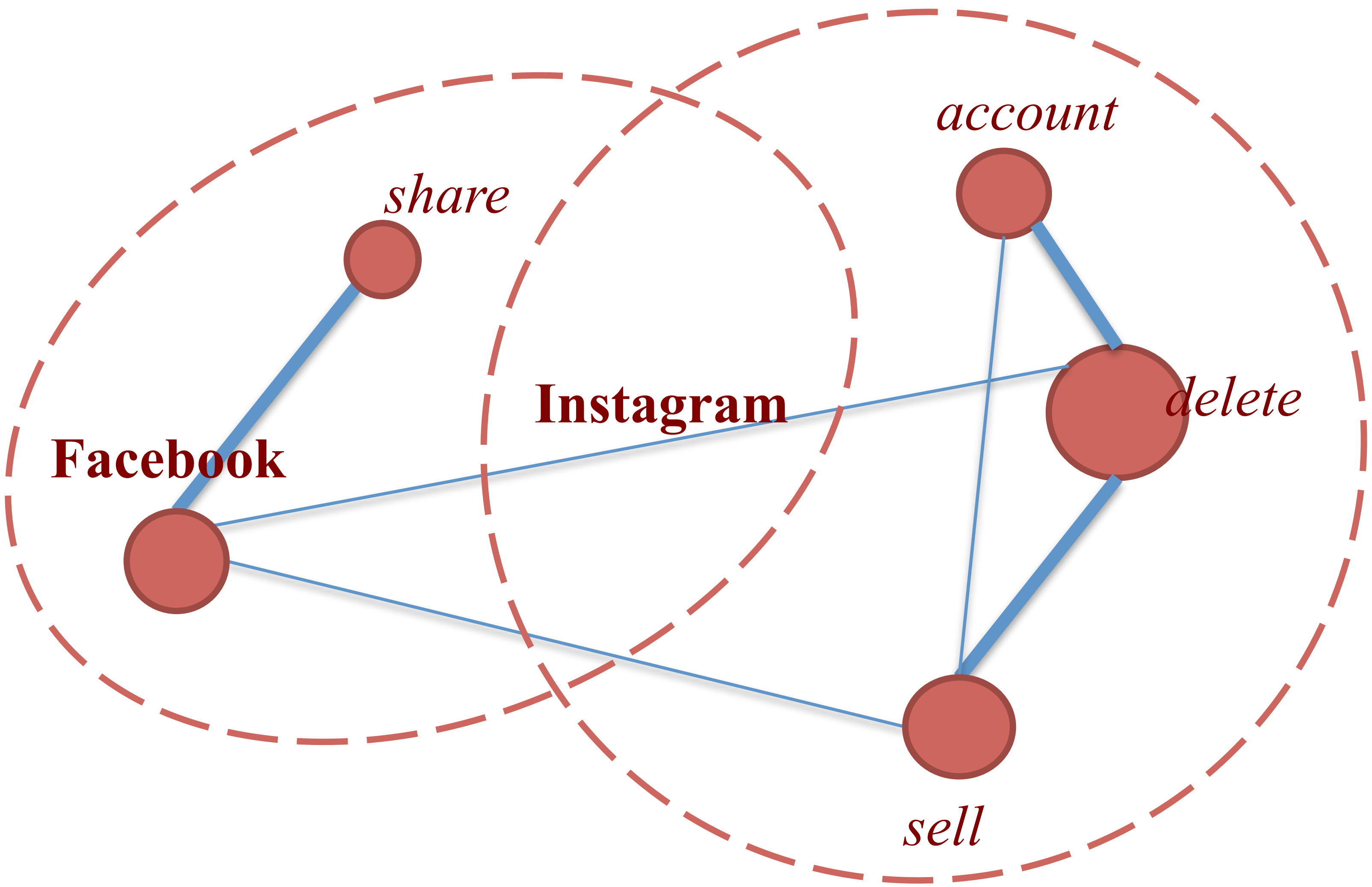
Graph Partitioning

- local graph partitioning by PageRank [Andersen et al., 2006] : a good partition of the graph can be obtained by separating high ranked vertices from low ranked vertices



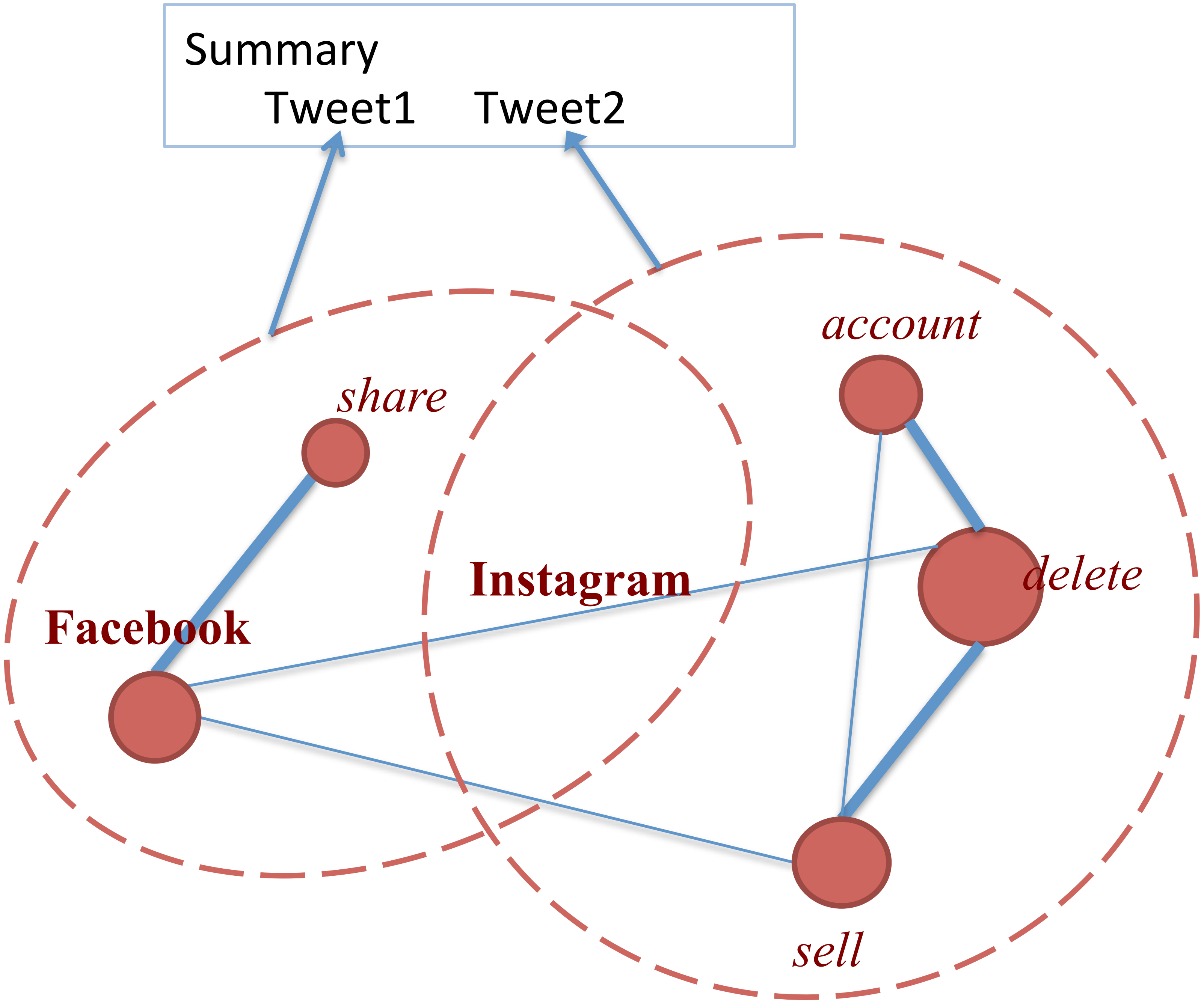
Wei Xu, Alan Ritter, Ralph Grishman.
"A Preliminary Study of Tweet Summarization using Information Extraction" in LASM (2014)

Graph Partitioning



Wei Xu, Alan Ritter, Ralph Grishman.
"A Preliminary Study of Tweet Summarization using Information Extraction" in LASM (2014)

Graph Partitioning



Wei Xu, Alan Ritter, Ralph Grishman.
"A Preliminary Study of Tweet Summarization using Information Extraction" in LASM (2014)

Example Event Graph

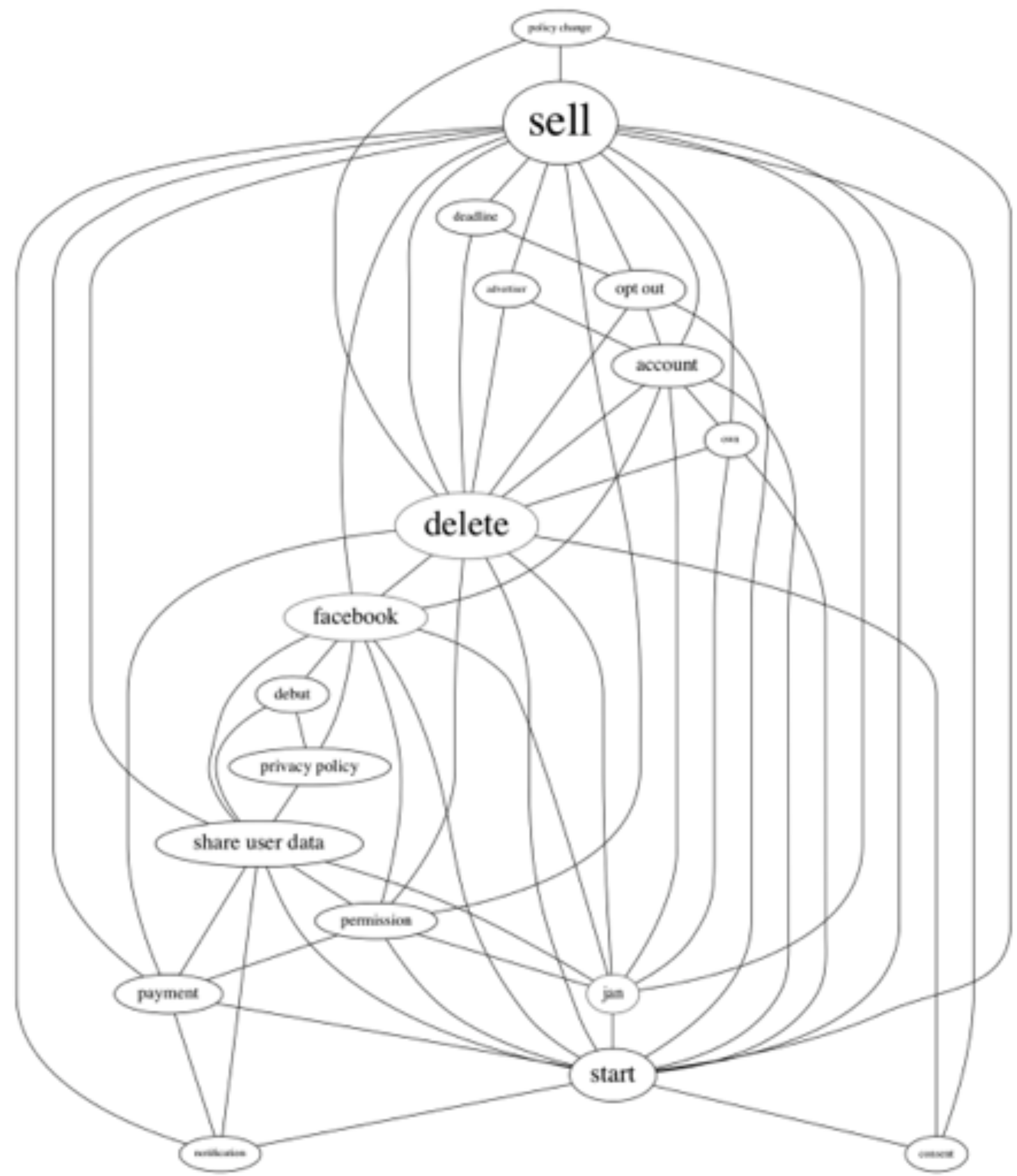


Figure 3: Event graph of 'Instagram - 1/16/2013', an example of event cluster with a single but complex focus

Wei Xu, Alan Ritter, Ralph Grishman.
"A Preliminary Study of Tweet Summarization using Information Extraction" in LASM (2014)

Example Summary

Instagram 1/16/2013	EventRank (Flexible)	- So Instagram can sell your pictures to advertisers without u knowing starting January 16th I'm bout to delete my instagram ! - Instagram debuts new privacy policy , set to share user data with Facebook beginning January 16
	SumBasic	- Instagram will have the rights to sell your photos to Advertisers as of jan 16 - Over for Instagram on January 16th - Instagram says it now has the right to sell your photos unless you delete your account by January 16th http://t.co/tsjic6yA

Wei Xu, Alan Ritter, Ralph Grishman.
"A Preliminary Study of Tweet Summarization using Information Extraction" in LASM (2014)

Example Event Graph

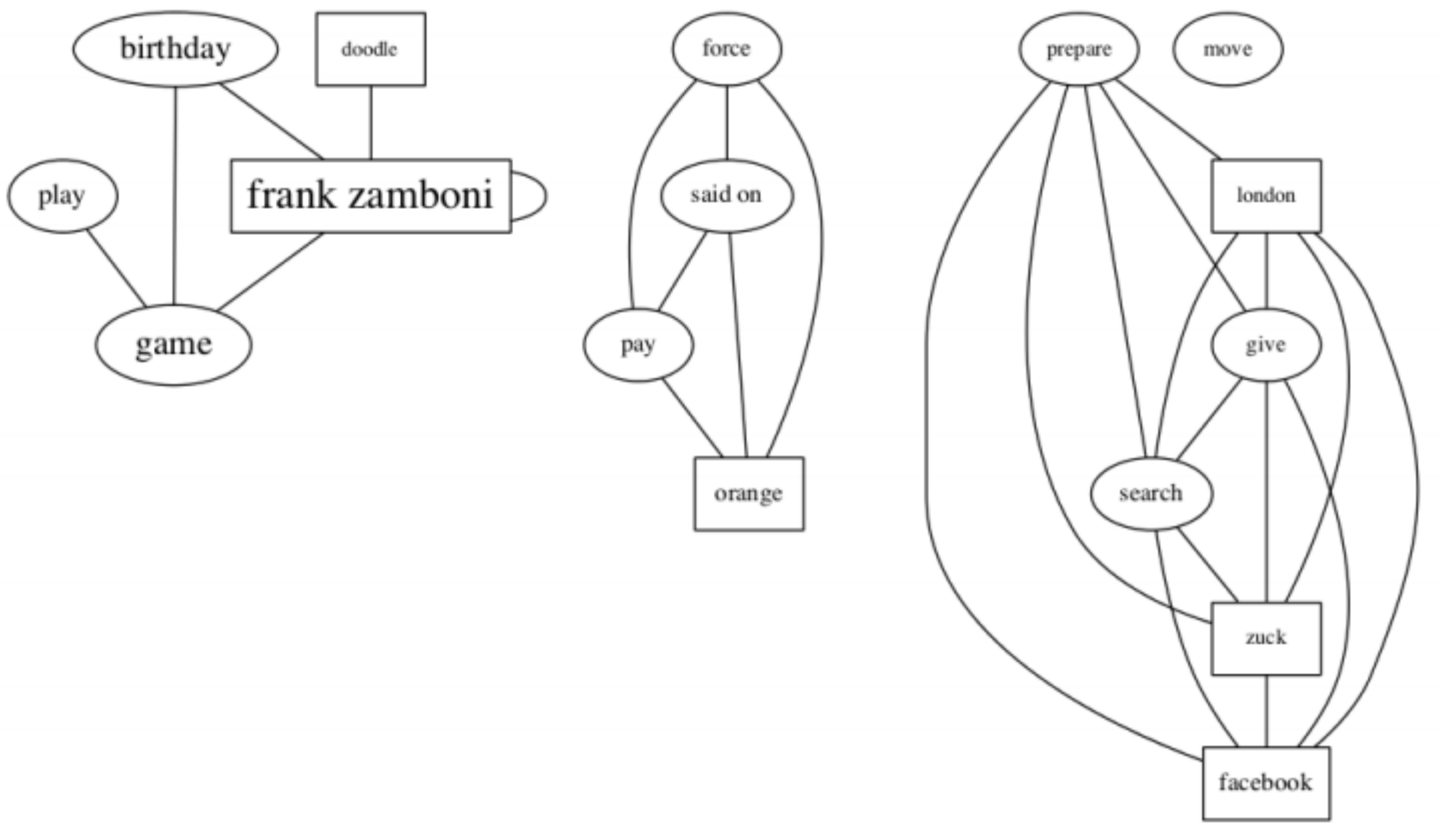


Figure 2: Event graph of 'Google - 1/16/2013', an example of event cluster with multiple focuses

Wei Xu, Alan Ritter, Ralph Grishman.
"A Preliminary Study of Tweet Summarization using Information Extraction" in LASM (2014)

Example Summary

Google 1/16/2013	EventRank (Flexible)	- Google 's home page is a Zamboni game in celebration of Frank Zamboni 's birthday January 16 #GameOn - Today social , Tomorrow Google ! Facebook Has Publicly Redefined Itself As A Search Company http://t.co/dAevB2V0 via @sai - Orange says has it has forced Google to pay for traffic . The Head of the Orange said on Wednesday it had ... http://t.co/dOqAHhWi
	SumBasic	- Tomorrow's Google doodle is going to be a Zamboni! I may have to take a vacation day. - the game on google today reminds me of hockey #tooexcited #saturday - The fact that I was soooo involved in that google doodle game says something about this Wednesday #TGIW You should try it!

Wei Xu, Alan Ritter, Ralph Grishman.
"A Preliminary Study of Tweet Summarization using Information Extraction" in LASM (2014)

[Recap]Research Questions

- What is the perfect length of multi-tweet summary?
- Will IE help summarization on Twitter?
  - noisy text: performance of IE?
  - short context: still need in-depth event analysis?
  - redundant: is word enough?

Wei Xu, Alan Ritter, Ralph Grishman.
"A Preliminary Study of Tweet Summarization using Information Extraction" in LASM (2014)

Research Questions

- What is the perfect length of multi-tweet summary?
  variable length

- Will IE help summarization on Twitter?
  - noisy text: performance of IE?
    summary is more readable and newsworthy
  - short context: still need in-depth event analysis?
    self-contained (no coref.) → better event graph
  - redundant: is word enough?
    unbalanced event graph → easier partitioning

Wei Xu, Alan Ritter, Ralph Grishman.
"A Preliminary Study of Tweet Summarization using Information Extraction" in LASM (2014)

Thank You!

Follow @cocoweixu

Instructor: Wei Xu

www.cis.upenn.edu/~xwe/

Course Website: socialmedia-class.org