A Web of Opinions

Sentiment Analysis in the Context of Online Communities

Bo Pang

YAHOO! RESEARCH

Sentiment analysis / Opinion mining

- What we do: computational treatment of opinion, sentiment, and subjectivity in text
  - SA at large: emotions, viewpoints, personal experience, ...
- Massive amount of opinion-oriented information online
  - review sites, forums, blogs, facebook status, tweets, ...
  - reviews, political discourse, ...
- Who cares?
  - researchers in natural language processing, information retrieval, data mining, ...

The importance of what strangers think

According to two surveys of more than 2000 users each

- 24% report using online reviews prior to paying for a service delivered offline
- between 73% and 87% online review readers report that reviews had a significant influence on their purchase;
- 32% have provided a rating; 30% have posted an online comment or a review

[comScore '07; Horrigan Pew survey '08]

How to **automatically** analyze such information?

The early days

single-document,

what is the polarity of the opinion: positive or negative?

[Das & Chen '01; Turney '02; Pang, Lee, & Vaithyanathan '02; ...]

The early days

single-document, bag-of-words representation

what is the polarity of the opinion: positive or negative?

and	Austen	beat	books	can’t
conceal	dig	Everytime	frenzy	from
her	I	Jane	madden	me
my	over	own	Prejudice	Pride
read	reader	shin-bone	skull	so
that	the	to	up	

[Das & Chen ’01; Turney ’02; Pang, Lee, & Vaithyanathan ’02; ...]

The early days

single-document, bag-of-words representation

what is the polarity of the opinion: positive or negative?

indicative terms

and	Austen	beat	books	can’t
conceal	dig	Everytime	frenzy	from
her	I	Jane	madden	me
my	over	own	Prejudice	Pride
read	reader	shin-bone	skull	so
that	the	to	up	

[Das & Chen ’01; Turney ’02; Pang, Lee, & Vaithyanathan ’02; ...]

The early days

single-document, bag-of-words representation

what is the polarity of the opinion: positive or negative?

and	Austen	beat	books	can't
conceal	dig	Everytime	frenzy	from
her	I			me
my	over			ride
read	reader			so
that	the			

indicative terms

machine learning algorithms discovered from movie review data: "Still,"

[Das & Chen '01; Turney '02; Pang, Lee, & Vaithyanathan '02; ...]

The early days

single-document, bag-of-words representation

what is the polarity of the opinion: positive or negative?

and	Austen	beat	books	can't
conceal	dig	Everytime	frenzy	from
her	I	Jane	madden	me
my	over	own	Prejudice	Pride
read	reader	shin-bone	skull	so
that	the	to	up	

what now???

[Das & Chen '01; Turney '02; Pang, Lee, & Vaithyanathan '02; ...]

opinion

Jane Austen's books madden me so that I can't conceal my frenzy from the reader. **Everytime I read 'Pride and Prejudice' I want to dig her up and beat her over the skull with her own shin-bone.**

opinion

Jane Austen's books madden me so that I can't conceal my frenzy from the reader. Everytime I read 'Pride and Prejudice' I want to dig her up and beat her over the skull with her own shin-bone.

- More sophisticated analysis (that I will not go into):
  - model the complexity of language: POS, parse trees,...
    - negation (not great != terrible), sarcasm, ..
  - fine-grained / sub-sentential analysis

opinion

Jane Austen's books madden me so that I can't conceal my frenzy from the reader. Everytime I read 'Pride and Prejudice' I want to dig her up and beat her over the skull with her own shin-bone.

- More sophisticated analysis (that I will not go into):
  - model the complexity of language: POS, parse trees,...
    - negation (not great != terrible), sarcasm, ..
  - fine-grained / sub-sentential analysis
- Beyond single documents

opinion about what

Jane Austen's books madden me so that I can't conceal my frenzy from the reader. **Everytime I read 'Pride and Prejudice' I want to dig her up and beat her over the skull with her own shin-bone.**

- More sophisticated analysis (that I will not go into):
  - model the complexity of language: POS, parse trees,...
    - negation (not great != terrible), sarcasm, ..
  - fine-grained / sub-sentential analysis
- Beyond single documents

whose opinion about what

Jane Austen's books madden me so that I can't conceal my frenzy from the reader. Everytime I read 'Pride and Prejudice' I want to dig her up and beat her over the skull with her own shin-bone.

- More sophisticated analysis (that I will not go into):
  - model the complexity of language: POS, parse trees,...
    - negation (not great != terrible), sarcasm, ..
  - fine-grained / sub-sentential analysis
- Beyond single documents

whose opinion about what

Jane Austen's books madden me so that I can't conceal my frenzy from the reader. Everytime I read 'Pride and Prejudice' I want to dig her up and beat her over the skull with her own shin-bone.

-- Mark Twain

- More sophisticated analysis (that I will not go into):
  - model the complexity of language: POS, parse trees,...
    - negation (not great != terrible), sarcasm, ..
  - fine-grained / sub-sentential analysis
- Beyond single documents

Example: "get out the vote"

[Thomas, Pang, & Lee '06]

Classify Congressional floor debates: support or oppose?

Example: "get out the vote"

[Thomas, Pang, & Lee '06]

Classify Congressional floor debates: support or oppose?

- **ind**ividual-document classifier (difficult)

Example: "get out the vote"

[Thomas, Pang, & Lee '06]

Classify Congressional floor debates: support or oppose?

- **ind**ividual-document classifier (difficult)
- agreement classifier provides the "**str**ength" of how likely two speakers agree with each other

Example: "get out the vote"

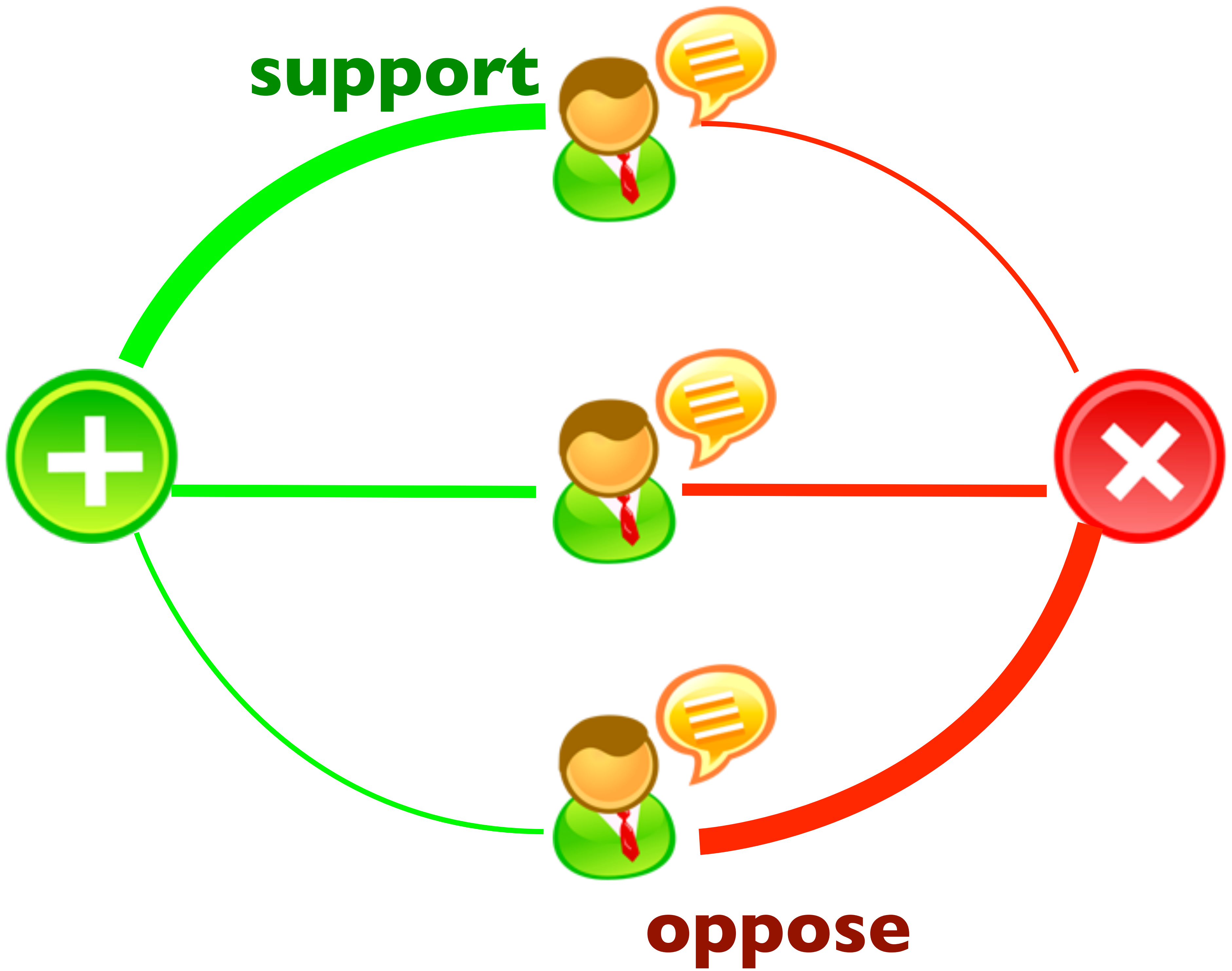
[Thomas, Pang, & Lee '06]

Classify Congressional floor debates: support or oppose?

- **ind**ividual-document classifier (difficult)
- agreement classifier provides the "**str**ength" of how likely two speakers agree with each other
- optimization problem: minimize

$$\sum_{s} ind(s, \overline{c}(s)) + \sum_{s,s' : \, c(s) \neq c(s')} \; \sum_{\ell \text{ between } s,s'} str(\ell)$$

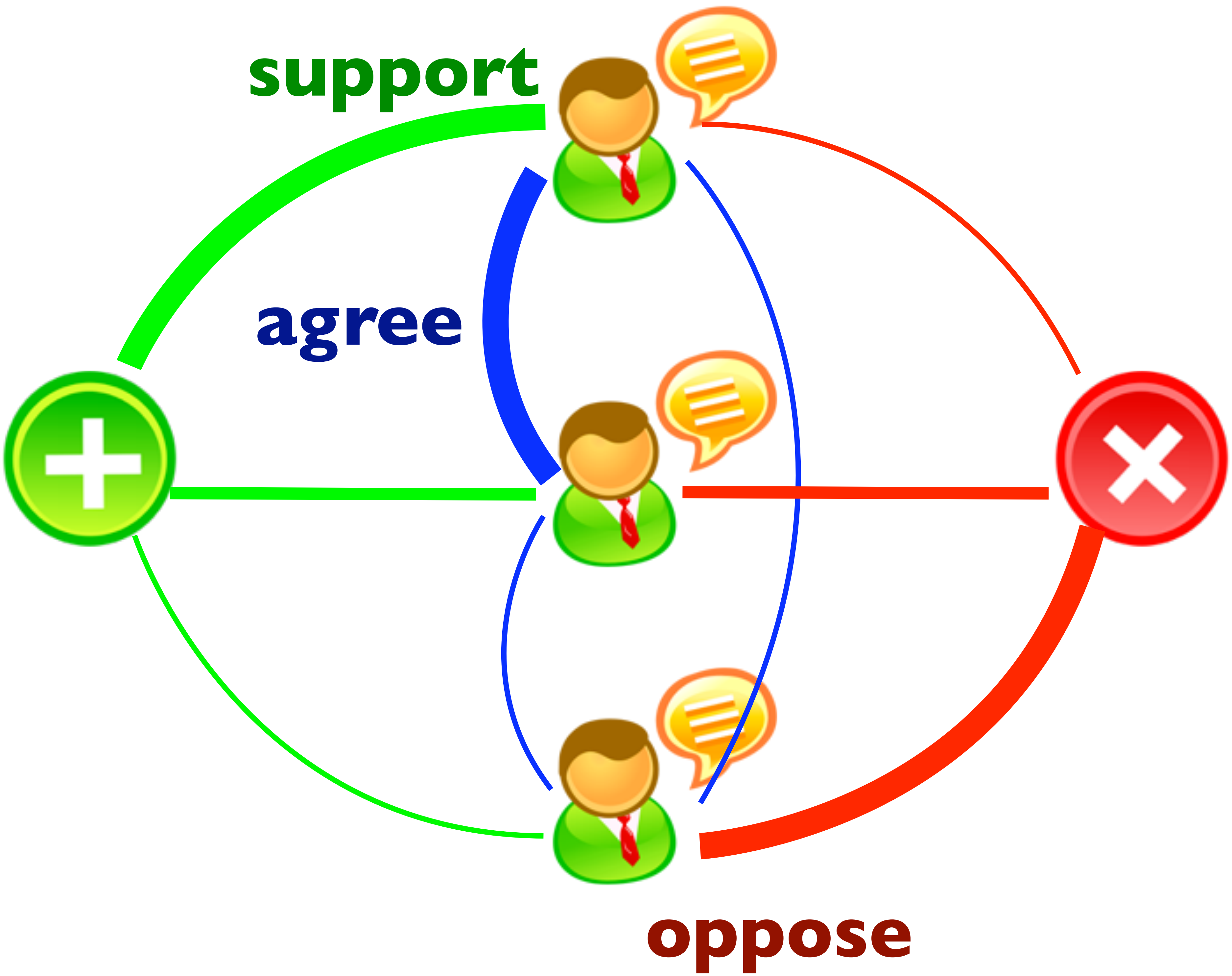
Graph-partitioning formulation



individual-document classifier

Held-out accuracy: 70%

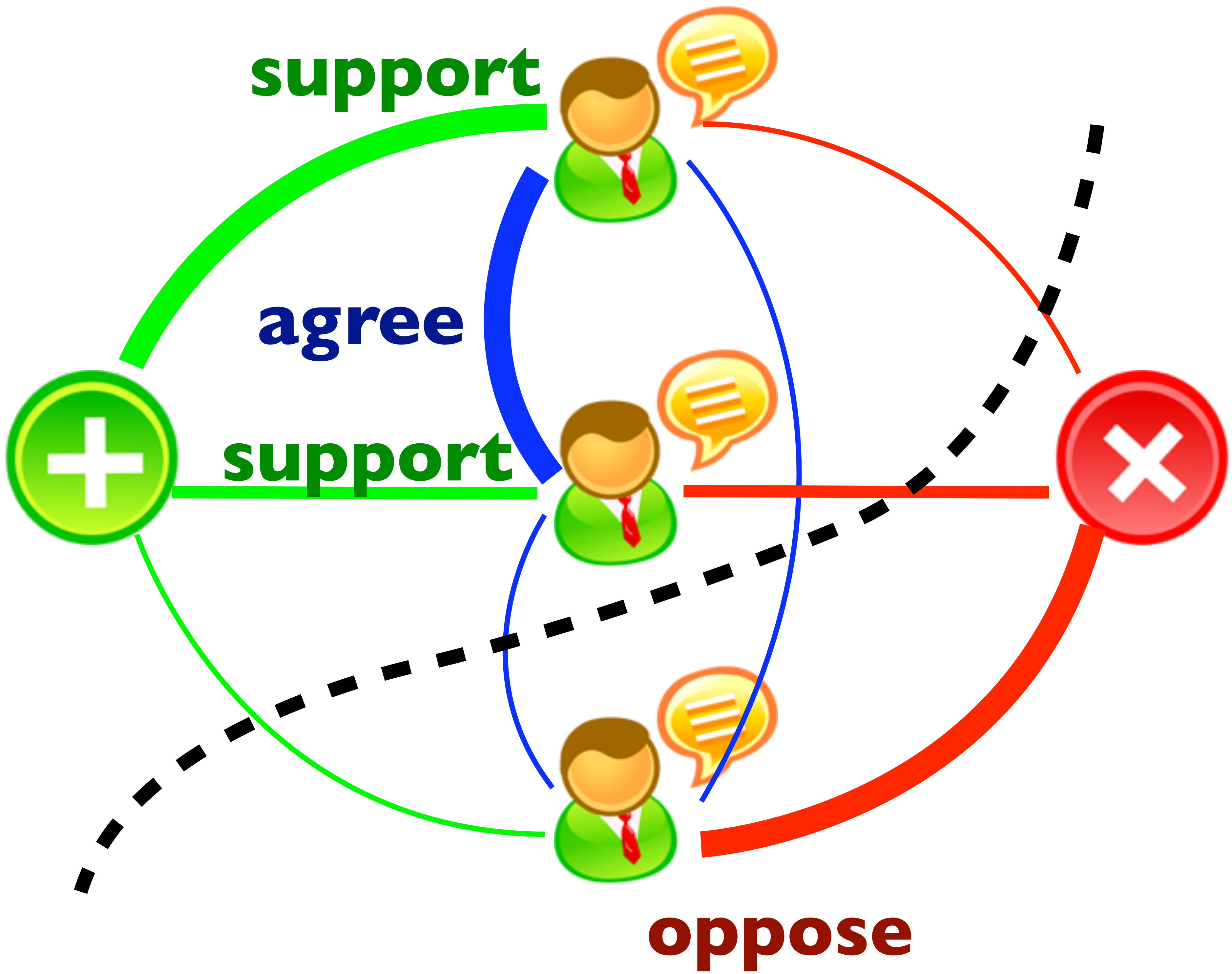
[Thomas, Pang, & Lee '06]



pair-wise agreement classifier

Held-out accuracy: 70%

[Thomas, Pang, & Lee ’06]

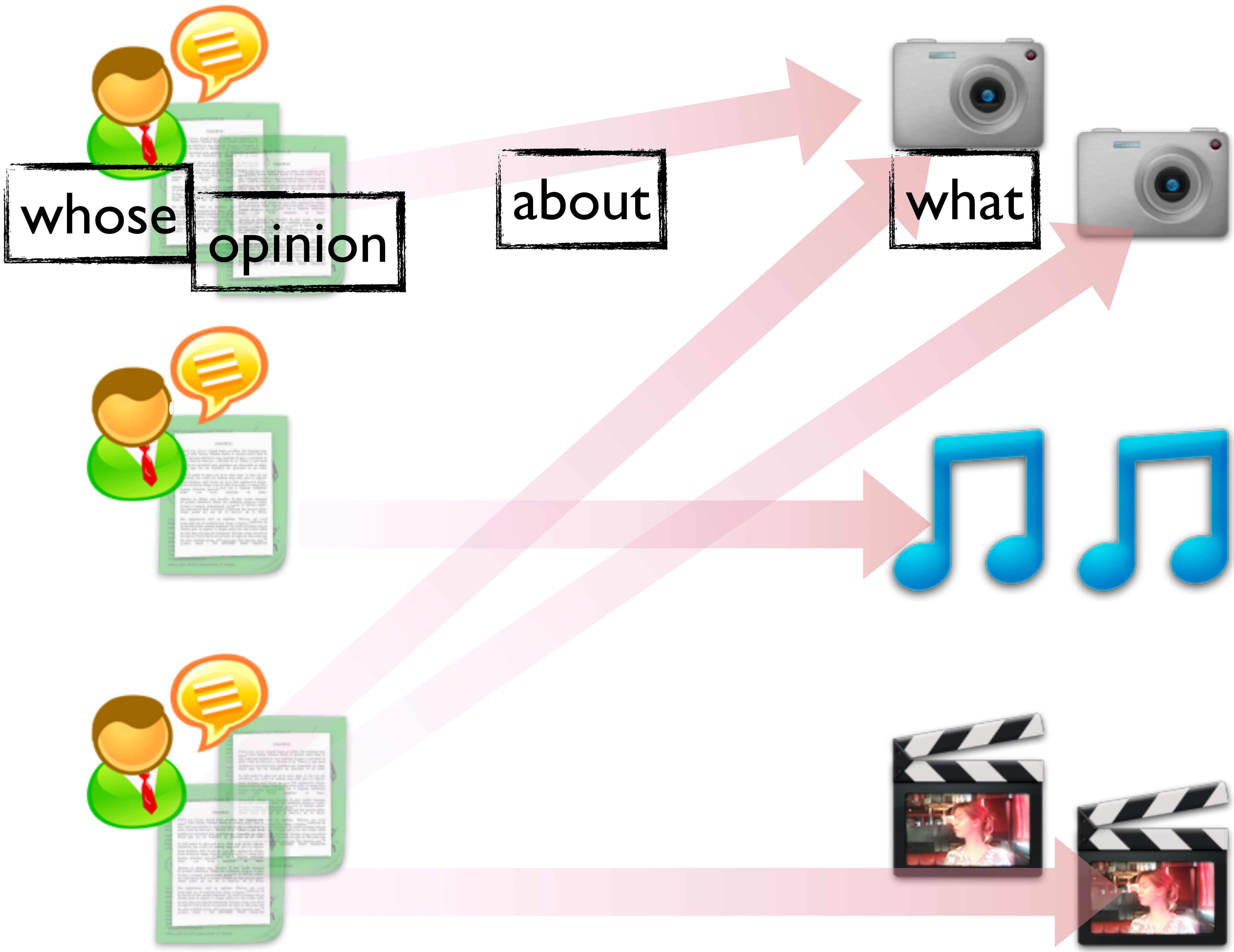


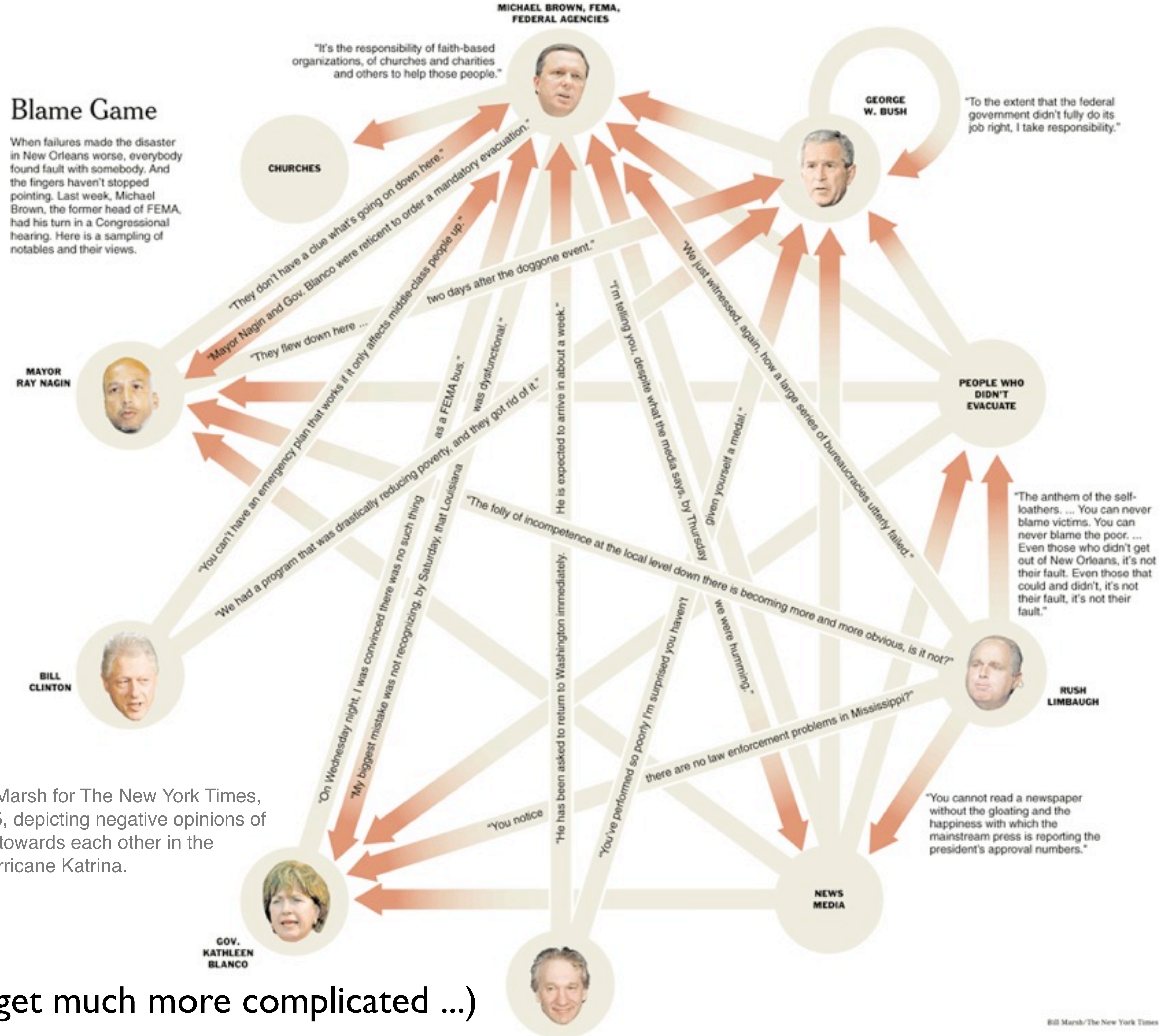
Held-out accuracy: 70%→76%

[Thomas, Pang, & Lee '06]

Opinions on the Web

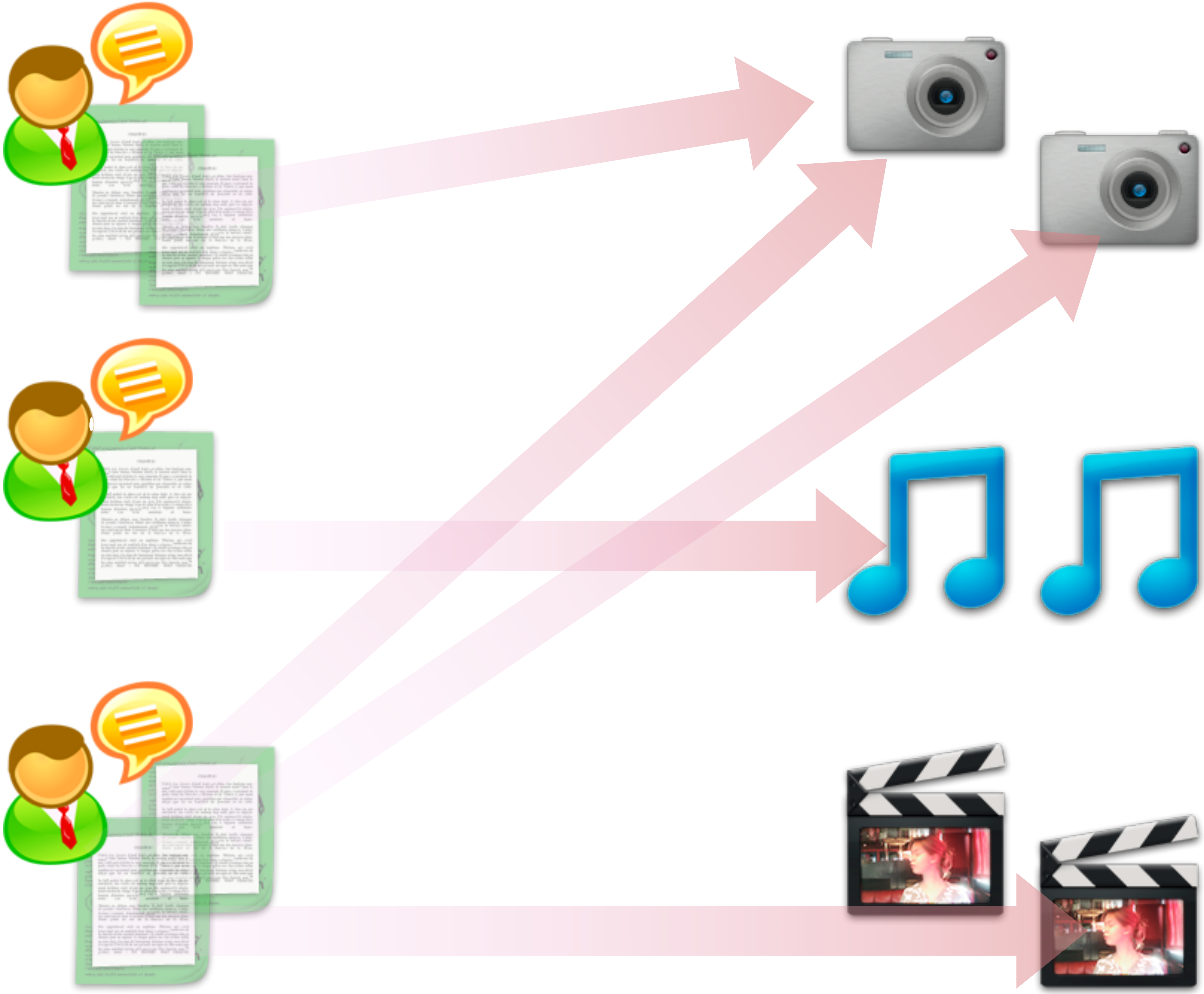
Opinions on the Web

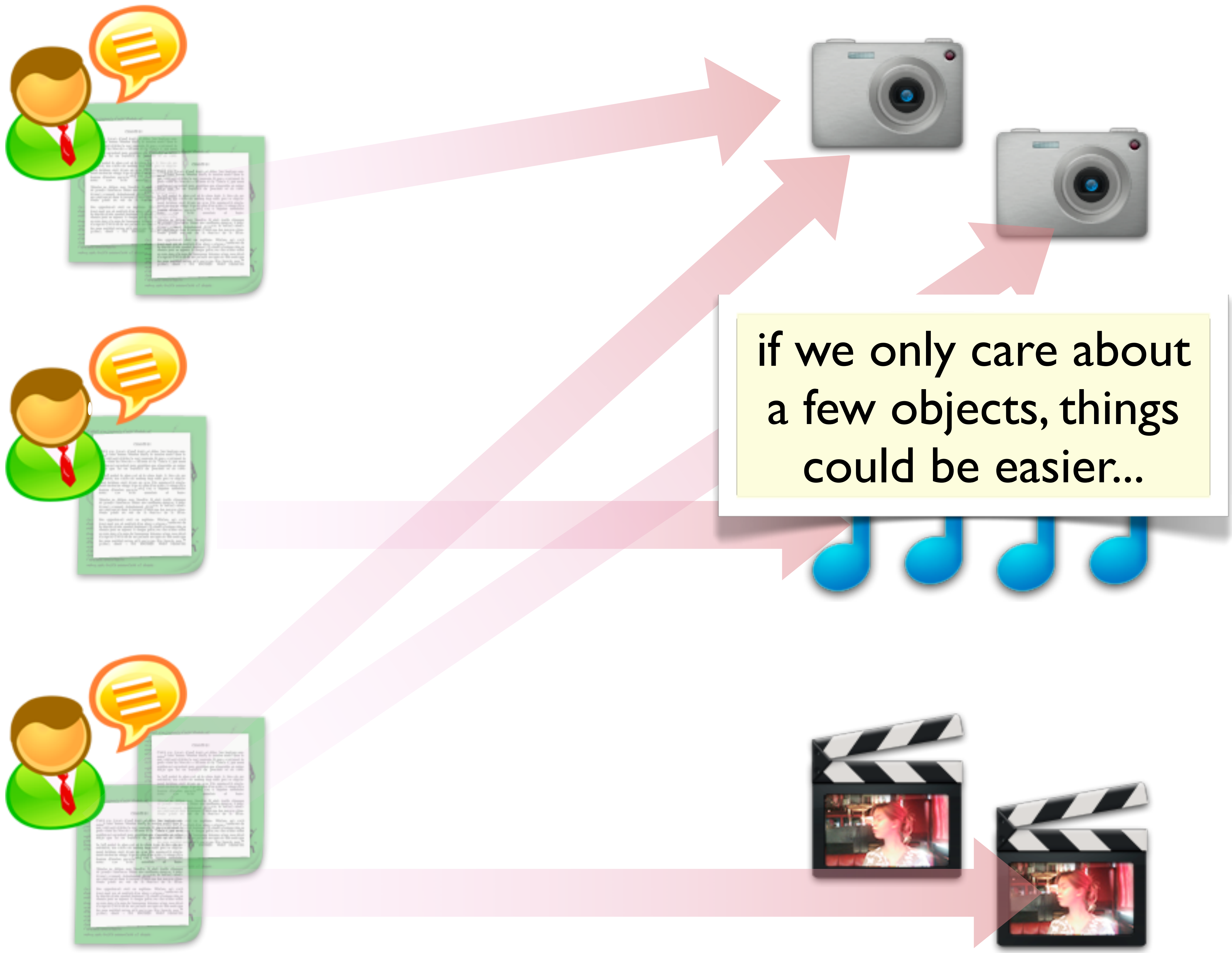

whose
opinion
about
what



Graphic by Bill Marsh for The New York Times, October 1, 2005, depicting negative opinions of various entities towards each other in the aftermath of Hurricane Katrina.

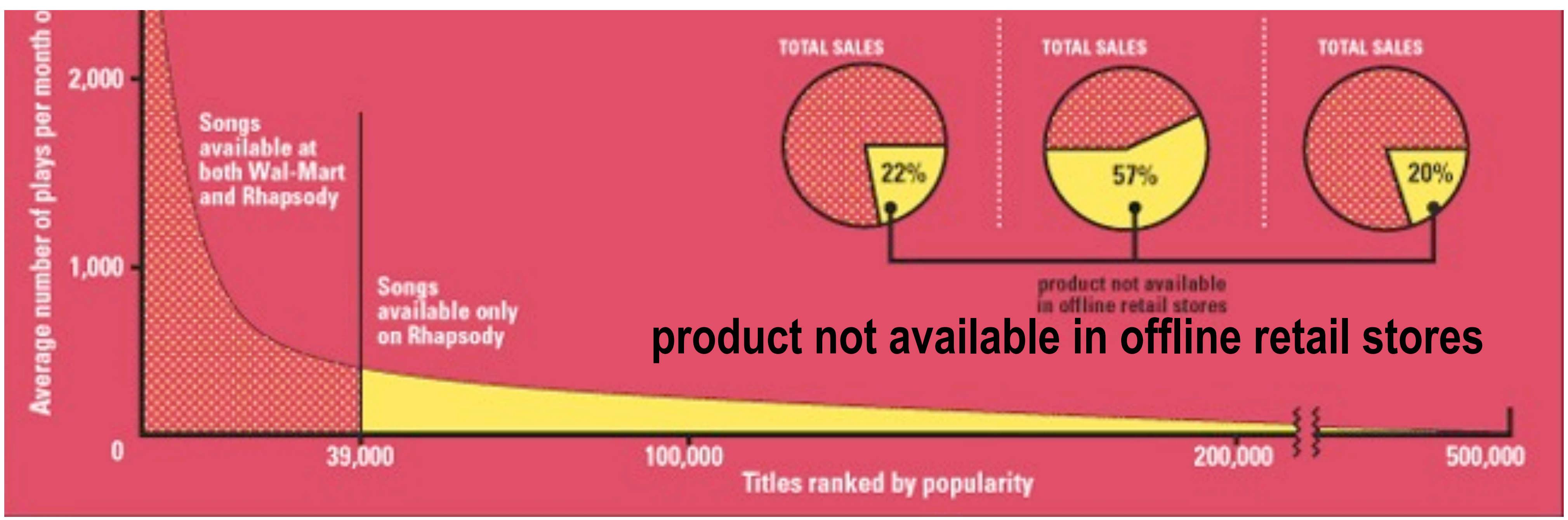
(it could get much more complicated ...)




if we only care about
a few objects, things
could be easier...

A quick digression: the long tail

- The tail is long: majority of products are "misses"
- In aggregate they account for a sizable fraction of total consumption; but much smaller compared to the head
- Who's consuming the tail? A few eccentric people?



Chris Anderson. The long tail. *Wired Magazine*, 2004

Ordinary people with extraordinary tastes

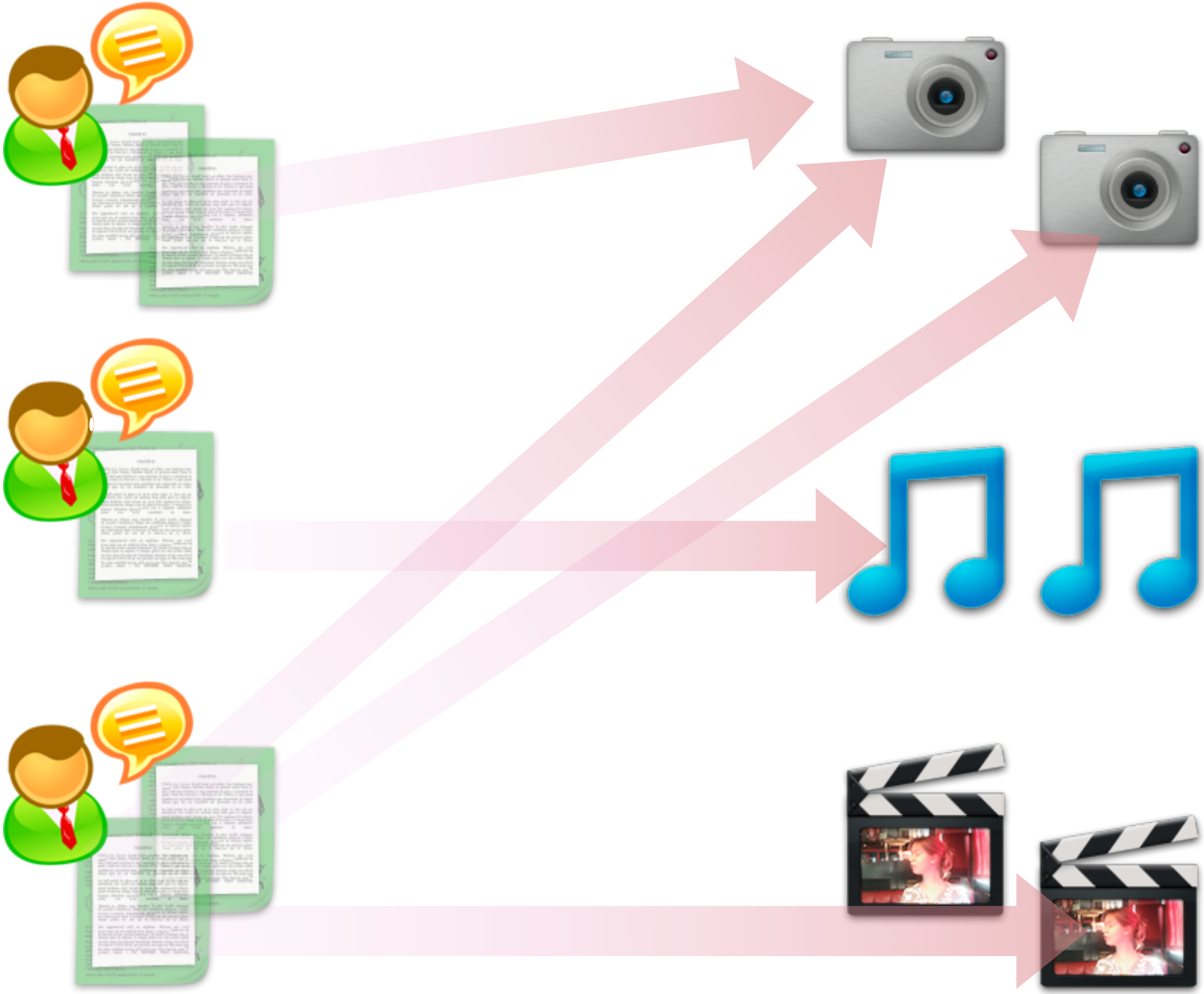
[Goel, Broder, Gabrilovich, & Pang '10]

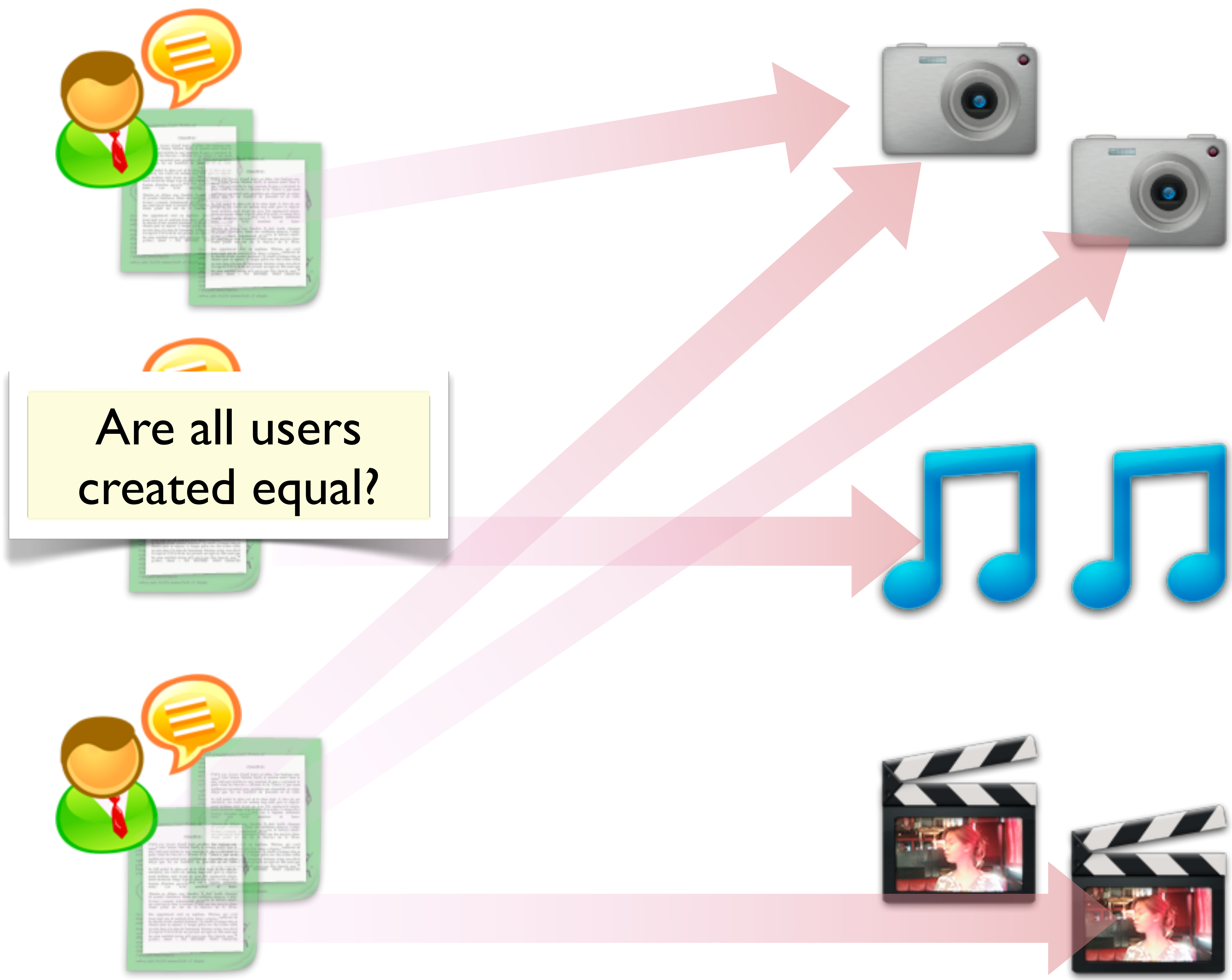
Nearly everyone is at least a bit eccentric

(movies, music, web search, and web browsing)

- 90% of Netflix users and 95% of Yahoo! Music users have consumed tail items
- 35% of Netflix users and 70% of Yahoo! Music users regularly do so

Understanding tail content is critical!




Are all users
created equal?

It sucks.

Overall Grade: F

by allisimlover (movies profile)
(Nov 22, 2004)
3 of 12 people found this review helpful

I loved this movie when I was little. Now I hate it! I've grown out of it. I might like it again,eventually. Burton rules!

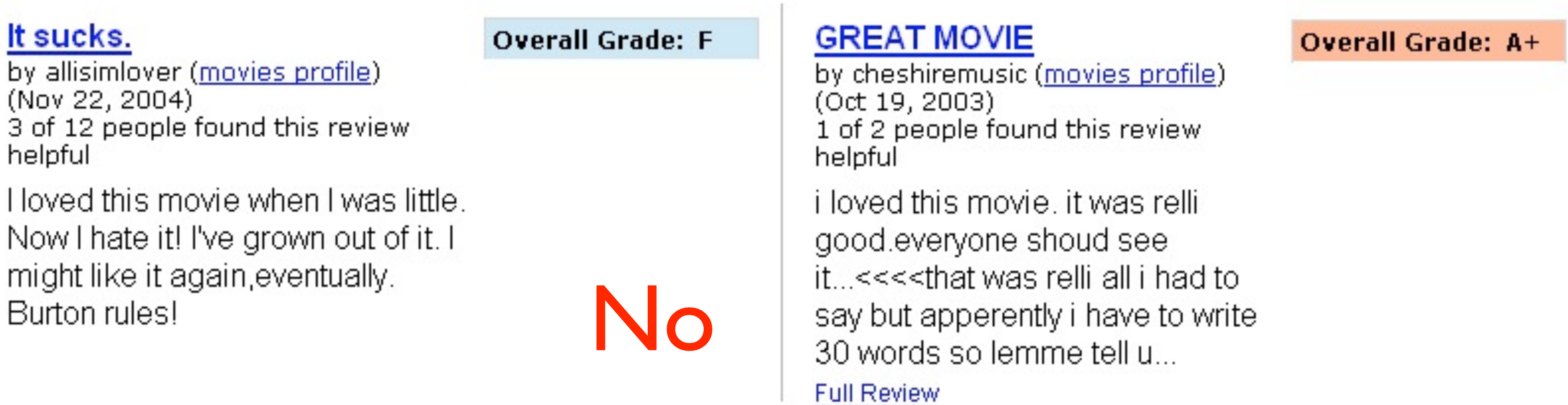
GREAT MOVIE

Overall Grade: A+

by cheshiremusic (movies profile)
(Oct 19, 2003)
1 of 2 people found this review helpful

i loved this movie. it was relli good.everyone shoud see it...<<<<that was relli all i had to say but apperently i have to write 30 words so lemme tell u...

Full Review



No

We need to figure out the "helpful" ones

- Utilizing user profiles
  [Danescu-Niculescu-Mizil, Kossinets, Kleinberg, & Lee '09]
  - using "real name"
- Utilizing social network [Lu, Tsaparas, Ntoulas, & Polanyi '10]
  - explicit "trust" relation given by reviewers
  - can we predict trust?
- Worse than bad: review spam [Jindal & Liu '08]

Yes, everyone counts!

From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series

Brendan O'Connor[†] **Ramnath Balasubramanyan**[†] **Bryan R. Routledge**[§] **Noah A. Smith**[†]
brenocon@cs.cmu.edu rbalasub@cs.cmu.edu routledge@cmu.edu nasmith@cs.cmu.edu

[†]School of Computer Science
Carnegie Mellon University

[§]Tepper School of Business
Carnegie Mellon University

Abstract

We connect measures of public opinion measured from polls with sentiment measured from text. We analyze several surveys on consumer confidence and political opinion over the 2008 to 2009 period, and find they correlate to sentiment word frequencies in contemporaneous Twitter messages. While our results vary across datasets, in several cases the correlations are as high as 80%, and capture important large-scale trends. The re-

statistics derived from extremely simple text analysis techniques are demonstrated to correlate with polling data on consumer confidence and political opinion, and can also predict future movements in the polls. We find that temporal smoothing is a critically important issue to support a successful model.

Data

- Sampling bias?

Anonymity: a double-edged sword

Anonymity: a double-edged sword

- Anonymity makes it more difficult to effectively consume opinions
    - it matters whose opinion it is

Anonymity: a double-edged sword

- Anonymity makes it more difficult to effectively consume opinions
  - it matters whose opinion it is
- Anonymity allows people to express their opinions more freely
  - this, in itself, can be good or bad

Anonymity: a double-edged sword

- Anonymity makes it more difficult to effectively consume opinions
  - it matters whose opinion it is
- Anonymity allows people to express their opinions more freely
  - this, in itself, can be good or bad
- Actually, hardly anyone is completely anonymous...

Anonymity in query logs

[Jones, Kumar, Pang, & Tomkins ’07, ’08]

- Given a sequence of queries issued by a given user + a simpler classifier, can we predict demographic info?
    - gender: accuracy 83.8%
    - age: avg. absolute error 7 years
- Vanity search
    - given a name, rank all the users who issued the name (modified tf/idf), 85% of the correct user rank at 1

Anonymity in query logs

[Jones, Kumar, Pang, & Tomkins '07, '08]

- Given a sequence of queries issued by a given user + a simpler classifier, can we predict demographic info?
  - gender: accuracy 83.8%

fanfiction, bridal, makeup, women's, knitting, hair, ecards, glitter, yoga, diet
nfl, poker, espn, ufc, railroad, prostate, football, golf, male, wrestling, compusa

  - age: avg. absolute error 7 years

- Vanity search
  - given a name, rank all the users who issued the name (modified tf/idf), 85% of the correct user rank at 1

Anonymity in query logs

[Jones, Kumar, Pang, & Tomkins '07, '08]

- Given a sequence of queries issued by a given user + a simpler classifier, can we predict demographic info?
  - gender: accuracy 83.8%

    fanfiction, bridal, makeup, women's, knitting, hair, ecards, glitter, yoga, diet
    nfl, poker, espn, ufc, railroad, prostate, football, golf, male, wrestling, compusa

  - age: avg. absolute error 7 years

    myspace, pregnancy, wikipedia, lyrics, quotes, apartments, torrent, baby, wedding, mall
    aarp, telephone, lottery, amazon.com, retirement, funeral, senior, mapquest, medicare,

- Vanity search
  - given a name, rank all the users who issued the name (modified tf/idf), 85% of the correct user rank at 1

Anonymity: privacy concerns vs. utility

- You are what you search for [Jones, Kumar, Pang, & Tomkins '07, '08]
- You are what you write [Novak, Raghavan, & Tomkins '04]
- You are who you connect to [Backstrom, Dwork, & Kleinberg '07]

  Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography

- You are the movies you watch [Narayanan & Shmatikov'08] ...

THE END (almost)

THE END (almost)

- The Web provides interesting raw data
  - easily reach out to opinions of millions of people
  - with intricate relationships

THE END (almost)

- The Web provides interesting raw data
  - easily reach out to opinions of millions of people
  - with intricate relationships
- Well, it's kind of messy...

THE END (almost)

- The Web provides interesting raw data
  - easily reach out to opinions of millions of people
  - with intricate relationships
- Well, it's kind of messy...
- But that's great -- challenge is opportunity!

More on applications, research directions, connections to other fields, ...

Opinion Mining and Sentiment Analysis
Bo Pang and Lillian Lee

www.cs.cornell.edu/home/llee/opinion-mining-sentiment-analysis-survey.html

135 pp, 330+ references, full pdf posted
Includes bibliographies, pointers to datasets, etc.