

It takes two to tango:  
Understanding the effects of language via  
“natural experiments”

Chenhao Tan  
Cornell University  
<https://chenhaot.com>

How can one “persuade” people, **using language?**

- Toward action (e.g., fighting in a war, voting, spreading the word, making your paper accepted)
- Toward different attitudes (e.g., angry, optimistic)

**Does language matter at all?**

# Rhetoric: dating from Ancient Greece

“Just because you do not take an interest in politics ... doesn't mean politics won't take an interest in you.”

*His speeches inspired Athenians to become the most powerful people in Greece. [<http://list25.com/25-speeches-that-changed-the-world/>]*



Pericles' Funeral Oration to Athenians during the Peloponnesian War (c. 430 BC)

## A long list of successful stories

Patrick Henry's "Give Me Liberty or Give Me Death"

The Gettysburg Address

Churchill's speeches during World War II

"Quit India" by Gandhi

...

Maybe these are only outliers,  
what about some “trivial” cases?

Debating about whether to buy orange juice for  
AI seminar at a faculty meeting.

Does the language still matter?

Maybe volume, or just a tan suit



## It is all about followers (Score:3, Interesting)

by mysterons (1472839) on Thursday May 15, 2014 @01:36PM (#47010441)

We did a study on predicting when a tweet would be retweeted (this paper cites us). The dominant factor is not what you write, but how many followers you have. Basically, a famous person can write anything and it will be retweeted. An unknown person can write the same tweet and it will be ignored.

Link to paper:

Sasa Petrovic, Miles Osborne and Victor Lavrenko. RT to win! Predicting Message Propagation in Twitter. ICWSM, Barcelona, Spain. July 2011. <http://homepages.inf.ed.ac.uk/...> [ed.ac.uk]

[Reply to This](#)

[Share](#)

Daniel Hopkins, SSRN 2013: “there is no evidence that groups targeted by specific frames [such as "death panels" in the health care debates] respond accordingly.”

# Lessons from science: experiments

Orange juice contains  
Vitamin C.



Representative group A

80% of PhDs like orange  
juice.



Representative group B



# Mobilizing Voter turnout

“How important is it to you to **be a voter** in the upcoming election?”



Representative group A

“How important is it to you to **vote** in the upcoming election?”



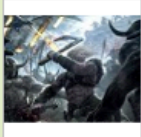
Representative group B

# Experiments are great, but they are difficult to scale

- Requires recruiting participants and asks for extra effort from participants
- Requires experiment designers to propose different wordings
- Lab can be different from real life

# Many online language+effect pairs

**The Instigator**




**Pro (for)**  
Lifeisgood

Abortion should be illegal in the United States

★ Add to My Favorites    🚩 Report this Debate    📎 Share with My Friends

Do you like this debate?   +25

**The Contender**



**Post Voting Period**

Vote Placed by NYCDiesel 4 years ago

	Lifeisgood	Xer	Tied	Points
Agreed with before the debate:	✓	-	-	0
Agreed with after the debate:	✓	-	-	0
Who had better conduct:	✓	-	-	1
Had better spelling and grammar:	✓	-	-	1
Made more convincing arguments:	✓	-	-	3
Used the most reliable sources:	✓	-	-	2
<b>Total points awarded:</b>	<b>7</b>	<b>0</b>		

Vote Placed by cahb 4 years ago

	Lifeisgood	Xer	Tied	Points
Agreed with before the debate:	-	-	✓	0
Agreed with after the debate:	-	✓	-	0 points
Who had better conduct:	-	✓	-	1 point
Had better spelling and grammar:	-	✓	-	1 point
Made more convincing arguments:	-	✓	-	3 points
Used the most reliable sources:	-	✓	-	2 points
<b>Total points awarded:</b>	<b>0</b>	<b>7</b>		



🗨️ [Request] Wine and apple sauce does not a dinner make.  
submitted 9 hours ago \* by WizardofStaz 🍕 got pizza'd  
4 comments share

🗨️ [Request] A pizza for starving students.  
submitted 9 hours ago \* by cuddlypotato  
2 comments share

🗨️ [Request] Columbus, OH. Celebration Pie!  
submitted 10 hours ago by noodlesdefyyou 🍕 pizza'd forward  
comment share

“How to Ask for a Favor: A Case Study on the Success of Altruistic Requests” Althoff, Danescu-Niculescu-Mizil, Jurafsky

# Effects of language on message propagation



**Barack Obama** ✓  
@BarackObama



Follow

Four more years. [pic.twitter.com/bAJE6Vom](http://pic.twitter.com/bAJE6Vom)

Reply Retweet Favorite More



RETWEETS  
**775,969**

FAVORITES  
**294,938**



8:16 PM - 6 Nov 2012

Flag media

“The effect of wording on message propagation: Topic- and author-controlled natural experiments on Twitter” Tan, Lee, Pang, ACL 2014.

# The same users post multiple tweets on the same topic

## Topic- and author-controlled pairs



**cactus\_music**  
@cactus\_music

I know at some point you've have been saved from hunger by our rolling food trucks friends. Let's help support them! [bit.ly/P6GYCq](http://bit.ly/P6GYCq)

7:59 PM - 15 Sep 2012

← ↻ ★



**cactus\_music**  
@cactus\_music

Food trucks are the epitome of small independently owned LOCAL businesses! Help keep them going! Sign the petition [bit.ly/P6GYCq](http://bit.ly/P6GYCq)

8:01 PM - 15 Sep 2012

← ↻ ★



**GeorgeMonbiot**  
@GeorgeMonbiot

read [@ameliagentleman](#)'s report today, then tell me Tories are no longer the nasty party: [guardian.co.uk/society/2012/o...](http://guardian.co.uk/society/2012/o...)

7:35 AM - 4 Oct 2012

← ↻ ★



**GeorgeMonbiot**  
@GeorgeMonbiot

Work capability tests: designed by bastards, performed by idiots. [guardian.co.uk/society/2012/o...](http://guardian.co.uk/society/2012/o...)

7:36 AM - 4 Oct 2012

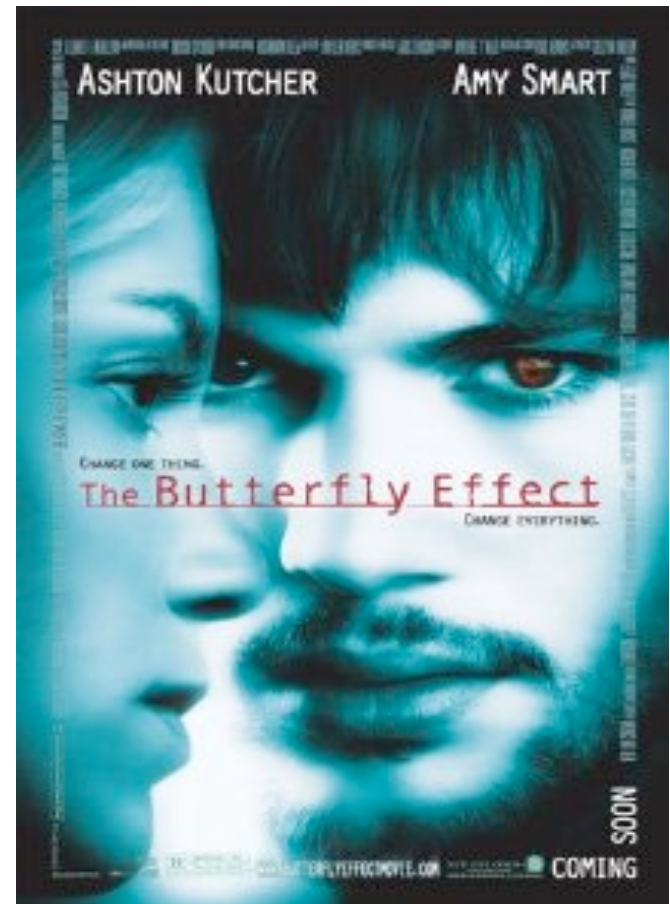
← ↻ ★



# Natural Experiment Paradigm

- *Same* speaker
- conveying the *same* info
- *Same* situation
- **Varies their wording**

and see the effects

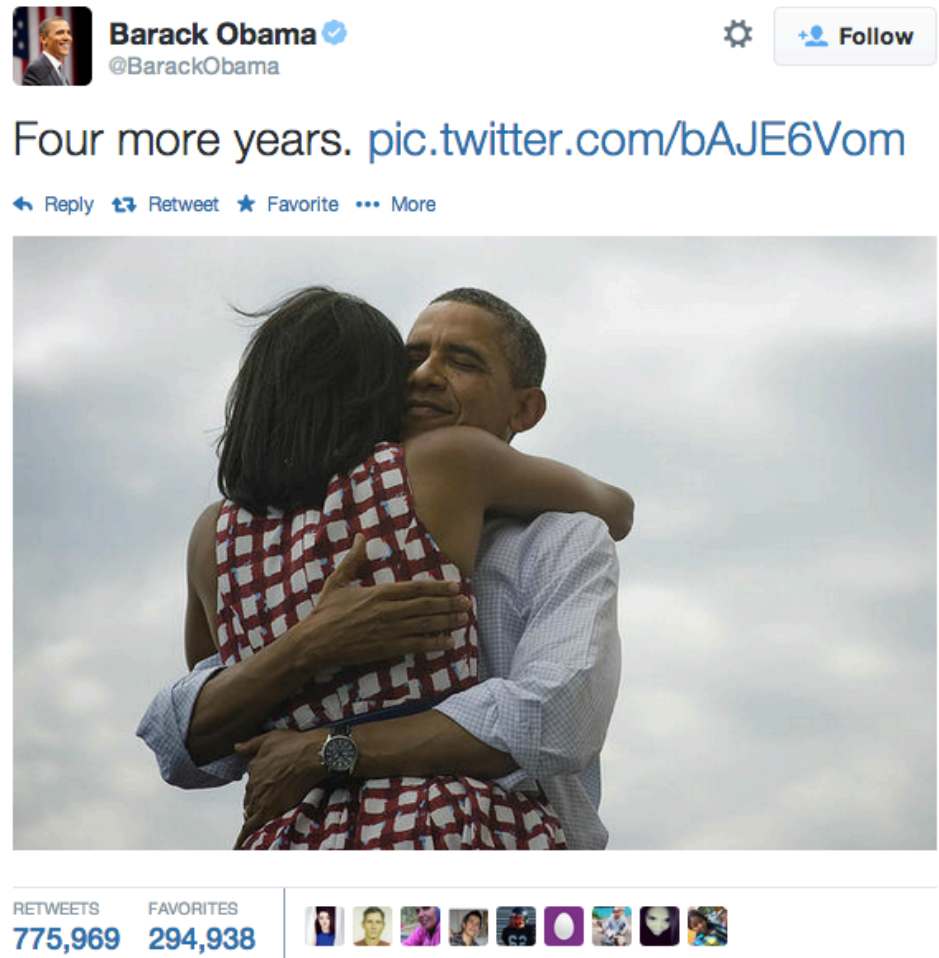


<http://www.imdb.com/title/tt0289879/>


# Existing literatures

**Important factors** [Milkman and Berger, 2012; Romero et al. 2013; Suh et al. 2010; etc]





- Characteristics of the author, author's social network
- Message topic
- Message timing



A screenshot of a tweet from Barack Obama (@BarackObama). The tweet text is "Four more years. [pic.twitter.com/bAJE6Vom](http://pic.twitter.com/bAJE6Vom)". Below the text are interaction icons for Reply, Retweet, Favorite, and More. The main image of the tweet shows Barack Obama embracing Michelle Obama from behind. They are outdoors, with a cloudy sky in the background. Michelle is wearing a red and white checkered dress. Barack is wearing a light blue shirt. Below the image, the tweet shows 775,969 retweets and 294,938 favorites, along with a row of profile pictures of users who interacted with the tweet.

**Barack Obama**   
@BarackObama

Four more years. [pic.twitter.com/bAJE6Vom](http://pic.twitter.com/bAJE6Vom)

 Reply  Retweet  Favorite  More

RETWEETS **775,969** FAVORITES **294,938**

8:16 PM - 6 Nov 2012

# How to get messages across more effectively?

- Find a good topic [Guerini et al. 2011]
- Become influential or find influential users to help spread [Kempe et al. 2003]
- Improve the quality of the content
  - Image [Isola et al. 2011]
  - **Wording**  
humor, informative, emphasize certain aspects



# Add topic- and author-control to understand the effects of language

- Author control
  - Obama vs. me
- Topic control
  - Presidential election vs. this talk

What if BarackObama had posted about re-election using a different wording?

e.g. “4 more years to prove that we can!”

# Topic- and author-controlled pairs are actually common!

- *2.4 Million* topic- and author-controlled tweet pairs
  - 1.77M differing in more than just spacing
  - 632K whose difference was only spacing

# More cleaning up is required for natural experiments!

- **Timing can matter** (thankfully, Twitter doesn't re-rank posts, but presents strictly in chronological order)
  - The first one may enjoy a first-mover advantage
  - The second one may be preferred as the updated one
- **Number of followers also has complicated effects**

# Use *identical pairs* to find an “ideal” setting

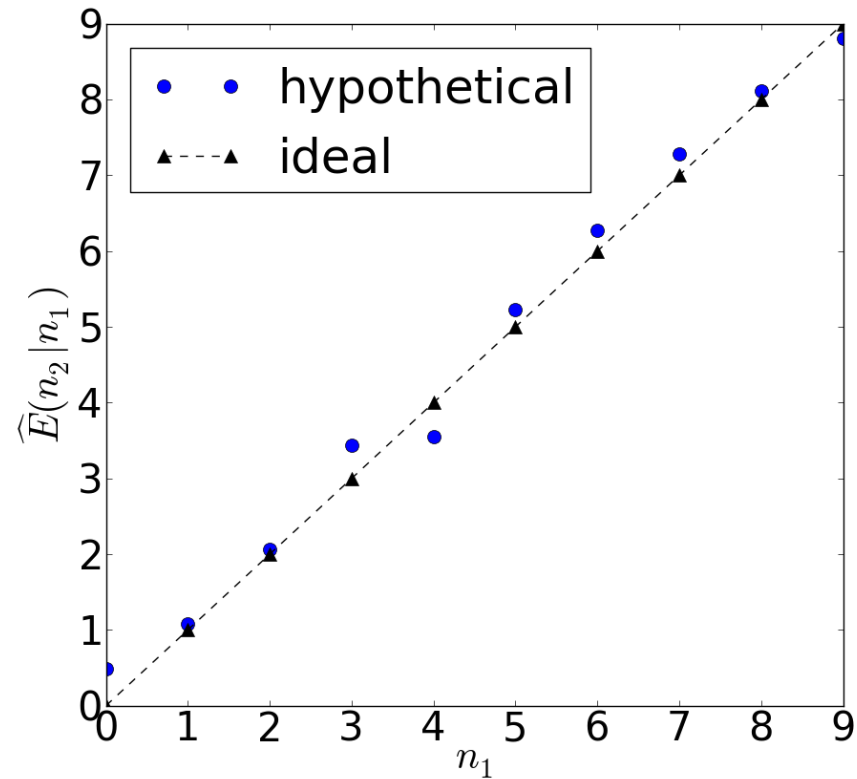
- Notation

- $n_1$  : number of retweets for the first tweet

- $n_2$  : number of retweets for the second tweet

- Difference between  $n_1$  and  $n_2$

$$D = \sum_{0 \leq n_1 < 10} |\hat{E}(n_2 | n_1) - n_1|$$

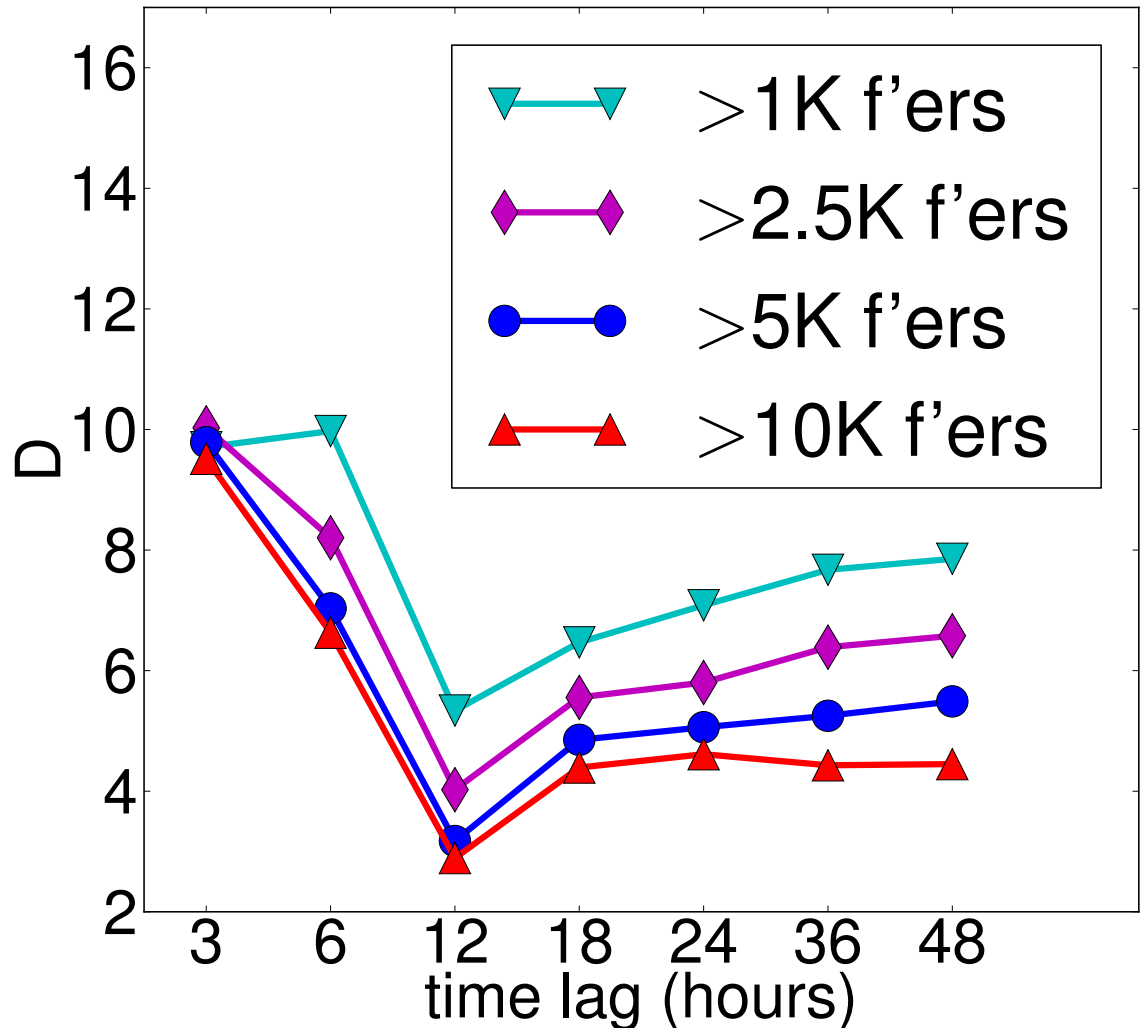


# Use *identical pairs* to find an “ideal” setting

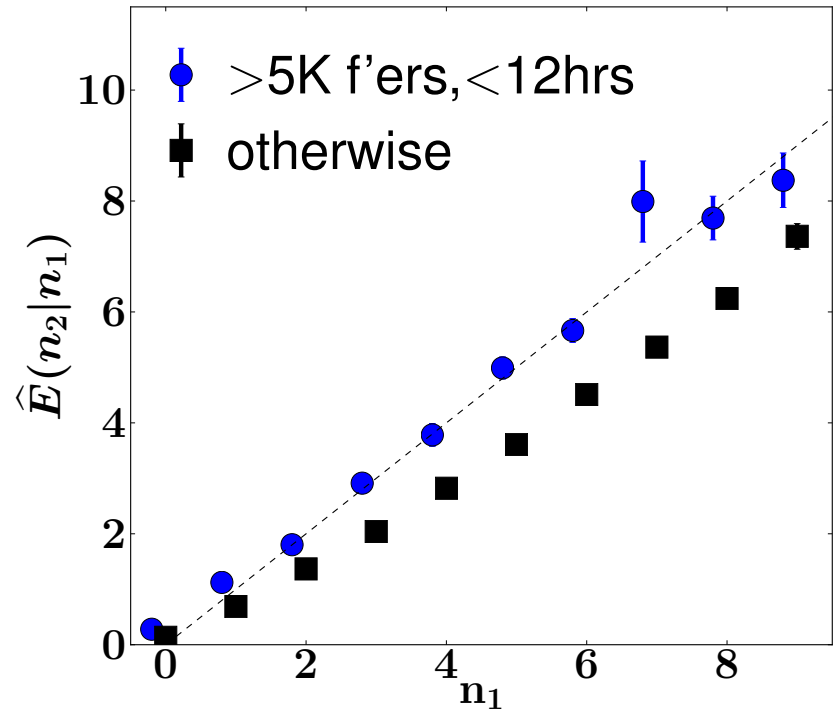
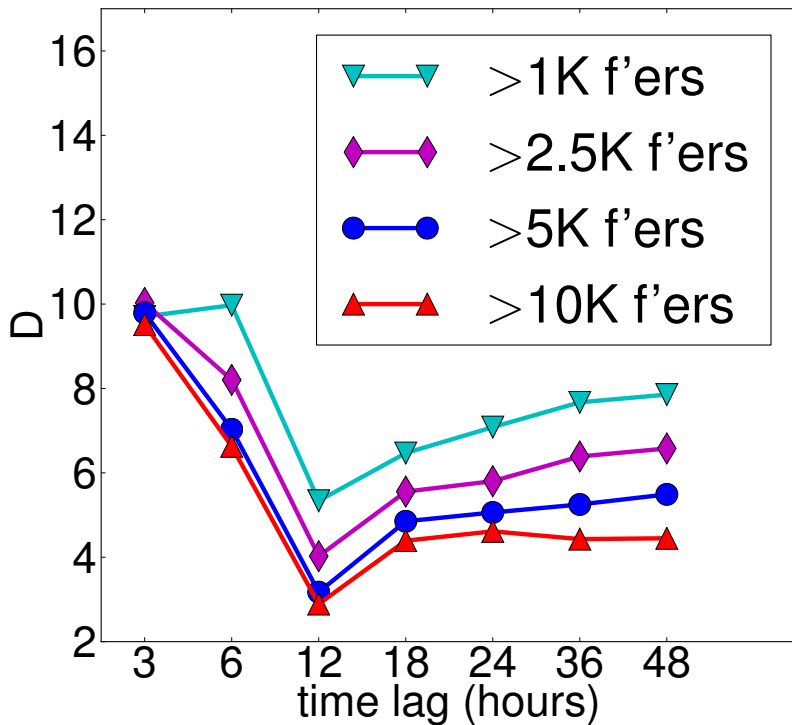
$$D = \sum_{0 \leq n_1 < 10} |\hat{E}(n_2 | n_1) - n_1|$$

As time lag increases,  $D$  decreases as we get more data and then increases

As number of followers increases,  $D$  decreases



The ideal setting found through *identical* pairs:  
users who have more than 5K followers  
two tweets are posted within 12 hours



# More filtering

- Ideal setting: >5K followers, <12 hours
- Non-trivial textual changes
  - Similarity below median to avoid typos, etc
- Significant changes in retweet numbers
  - Take top 5% and bottom 5% in terms of  $n_2 - n_1$
- Limit the number of pairs by an author to 50

This brings us 11K topic- and author- controlled pairs for natural experiments!

# Does wording matter?

Wording does not matter



Humans should not be able to tell which one in a pair was retweeted more

Humans can tell which one in a pair was retweeted more (accuracy > 50%)



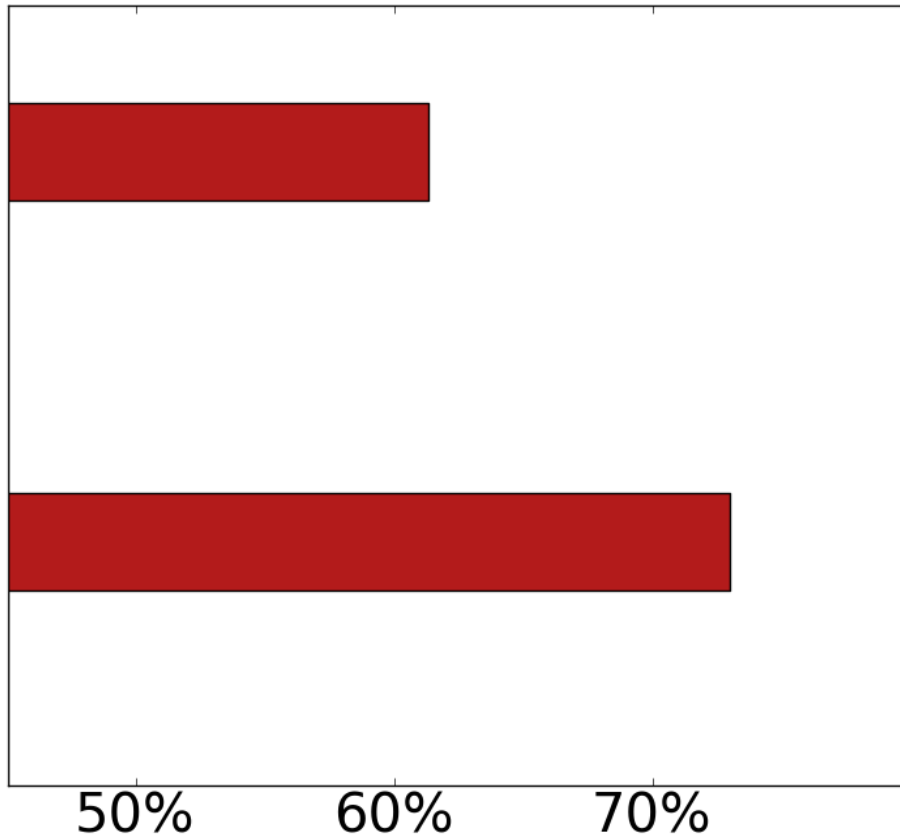
Wording matters!



# Can humans tell which tweet will be retweeted more?

- Randomly sample 100 pairs
- 20 pairs a task on Amazon Mechanical Turk
- 39 judgments for each pair

# Can humans tell which tweet will be retweeted more?



Average accuracy for each labeler: 61.3%

Accuracy of the majority label for each pair: 73%

# Predict which tweet will be retweeted more within a pair

- Cross validation experiments: 11K topic- and author-controlled pairs (5-fold cross validation)
- *Heldout* experiments: 1.8K topic- and author-controlled pairs from a different group of users that have never been used  
(Only used once, 6 days before submission!)

# Predict which tweet will be retweeted more within a pair

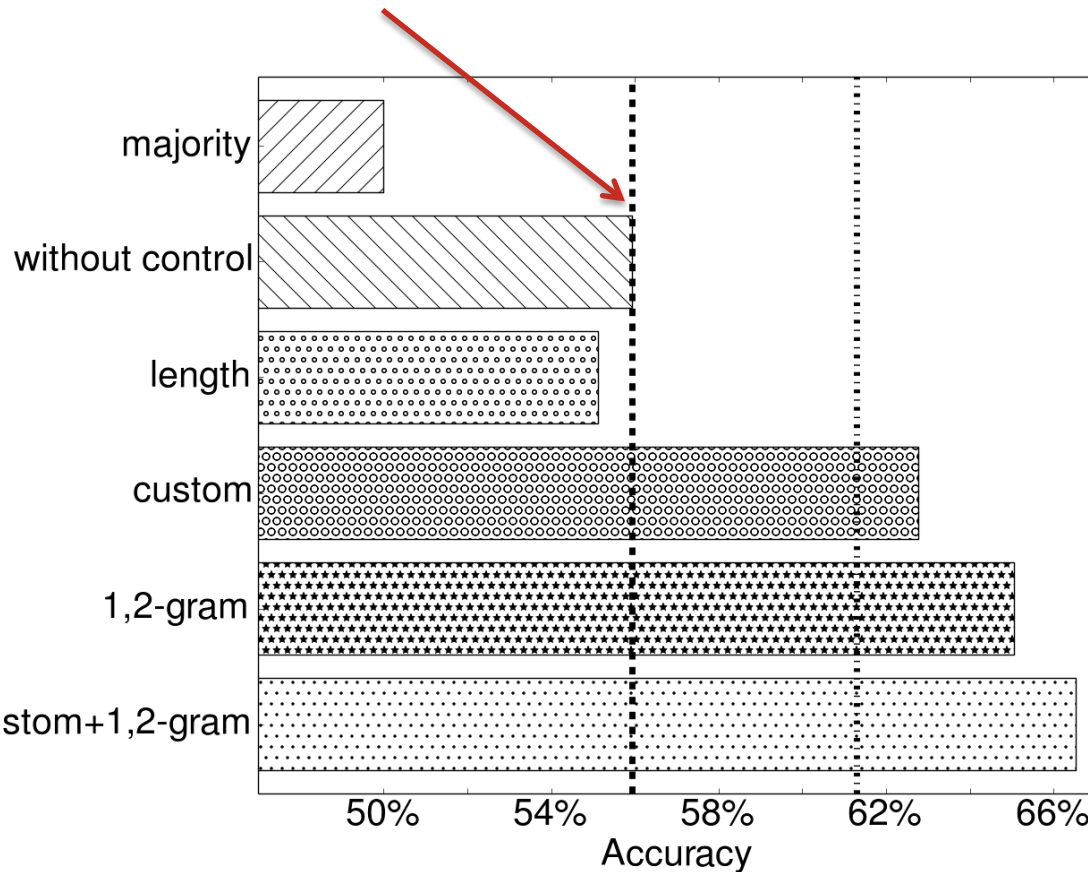
- Features
  - Custom features that we proposed: lexicons, informativeness, language model features, etc (39 features)
  - Bag of words: unigram+bigram (7K features)
- Approach
  - Take the difference between features for two tweets in a pair after linear normalization
  - Logistic regression

# Predict which tweet will be retweeted more within a pair

- A strong baseline that takes only ONE
  - A classifier to distinguish 10K most retweeted unpaired tweets from 10K least retweeted unpaired tweets
  - Use bag-of-words features, [number of followers and timing]
  - Cross validation accuracy 98.8%

# Cross-validation performance: is control necessary?

Accuracy without control

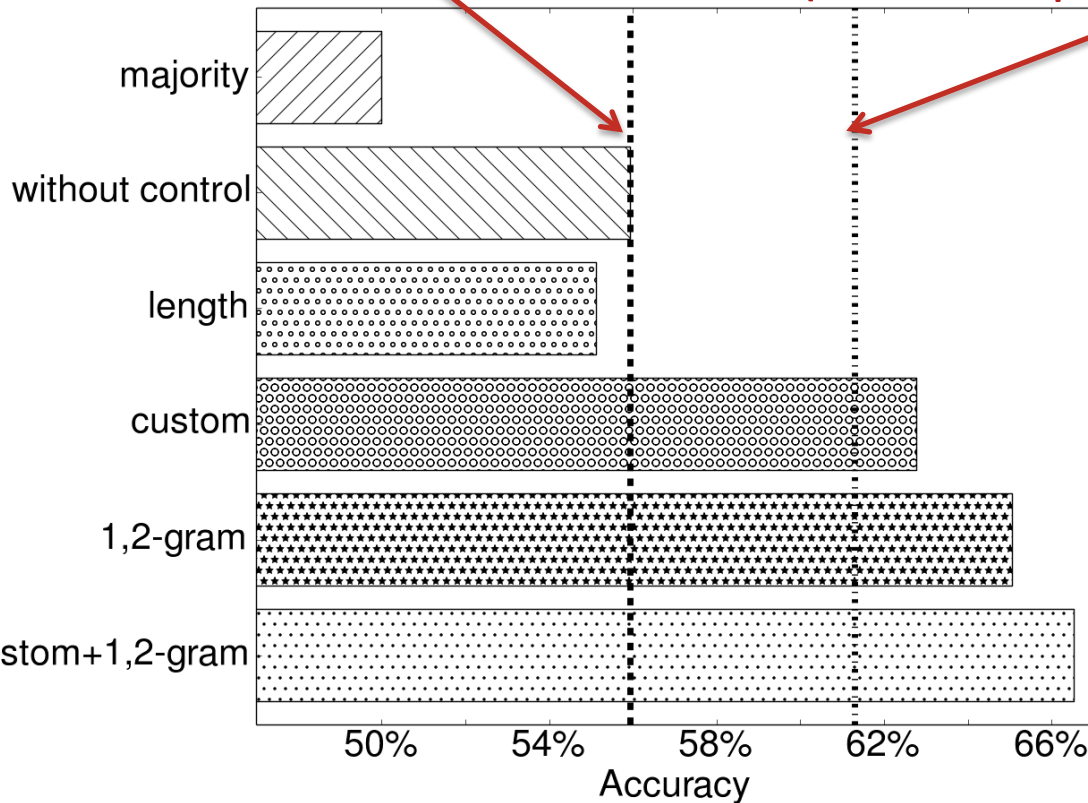


- Best method outperforms the baseline by more than 10%

# Cross-validation performance

Accuracy without control

Average human accuracy  
(on a sample of 100 pairs)

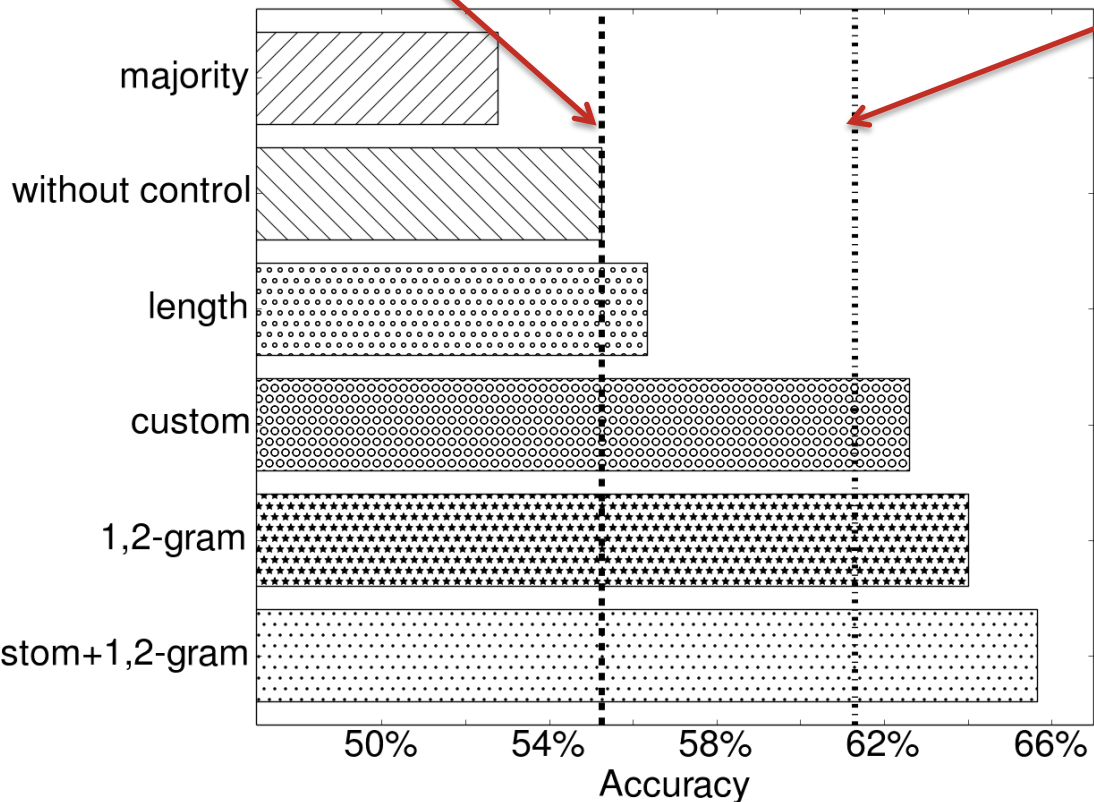


- Best method outperforms the baseline by more than 10%
- Custom does pretty well by itself, and outperforms average human accuracy
- Adding custom improves bag-of-words

# Fortunately, same results hold in heldout data

Accuracy without control

Average human accuracy  
(on a sample of 100 pairs)



- Best method outperforms the baseline by more than 10%
- Custom does pretty well by itself, and outperforms average human accuracy
- Adding custom improves bag-of-words



# Should we conform to community norm?

- Train language models using non-paired tweets
- Compute unigram, bigram language model score
  - higher score = closer to twitter language
- Test whether more retweeted tweets have a larger score

# Be like the community (conformity)

- Train language models using non-paired tweets
- Compute unigram, bigram language model score
  - higher score = closer to twitter language
- Test whether more retweeted tweets have a larger score

	Effective?
Twitter unigram language model	$p < 0.001$
Twitter bigram language model	$p < 0.001$

# Should we maintain personal style?

- Train language models using history of each person
- Compute unigram, bigram language model score  
higher score = closer to personal history
- Test whether more retweeted tweets have a larger score

# Be true to yourself

- Train language models using history of each person
- Compute unigram, bigram language model score  
higher score = closer to personal history
- Test whether more retweeted tweets have a larger score

	Effective?
Personal unigram language model	$p < 0.001$
Personal bigram language model	————

- Natural experiments show that language matters in message propagation!
- Controlling topics and authors can improve predictive performance significantly over an approach without control

Use similar paradigm to approach less  
studied problems:  
language strength

“A Corpus of Sentence-level Revisions in Academic Writing: A Step towards Understanding Statement Strength in Communication.” Tan and Lee, ACL 2014

# Example: Kunming Attack



The members of the Security Council (UN) condemned in the strongest terms the terrorist attack on March 1, 2014 in Kunming Train Station

[Chinese media] accused Western media of soft-pedaling the attack and failing to state clearly that it was an act of terrorism.” [The New York Times]

“Some Western media, including CNN, The Associated Press, The New York Times and The Washington Post, were mystifying, confusing, even to the point of sowing discord.”

‘Completely hypocritical and callous,’ [People’s daily]

In particular ...



..., the US embassy referred to this incident as the “terrible and senseless act of violence in Kunming”.

## **After Prodding, U.S. State Department Labels Kunming Attack ‘Terrorism’**

By DIDI KIRSTEN TATLOW MARCH 4, 2014 1:05 AM  19 Comments

A weibo user: “If you say that the Kunming attack is a ‘terrible and senseless act of violence’, then the 9/11 attack can be called a ‘regrettable traffic incident’”

# Understanding statement strength is important!



## EDUCATIONFORUM

EDUCATION

### Open Learning at a Distance: Lessons for Struggling MOOCs

Patrick McAndrew\* and Eileen Scanlon

Five education is changing how people think about learning online. The rise of Massive Open Online Courses (MOOCs) (1) shows that large numbers of learners can be reached. It also raises questions as to how effectively they support learning (2). There is a timeliness in the introduction of MOOCs, reflecting the right combination of online systems, interest from good teachers in reaching more learners, and banks of digital resources, predicted as a "perfect storm of innovation" (3). However, learning at scale, at a distance, is not a new phenomenon. Seeing MOOCs narrowly as a technology that expands access to in-classroom teaching can miss opportunities. Drawing on decades of lessons learned, we set out aims to help spur innovation in science education. Education based on gathering people together into a physical location is limited to those who can afford it and who make it past the filters that attenuate participation in higher levels of education. Those filters are inevitable on cost grounds to meet global needs "would require four major campus universities... to open every week" (4). The arrival of MOOCs highlights that there are alternatives. With courses enrolling over 100,000 students, MOOCs can reach students who have breaks in study, change where they study, mix study with work, and take at least part of their study online. Such students are now the majority, forming more than 70% are now in U.S. post-secondary education (5).

**Recommendations for Open Learning**  
We ought not behave as if learning at scale is unexplored territory and that there is no previous experience in being massive, open, or even online, upon which to build. Distance educationists, such as The Open University (OU) established in Britain more than 40 years ago, from their inception, ran courses for thousands of learners, accepted open entry, and led the move into online methods of teaching and learning. In each case, they provide lessons likely to apply in the new context of MOOCs.

**Build on distance-learning pedagogy:** Some of the steps taken toward "massive" classes simply follow the observation that a lecture presented to a few hundred students can be viewed by many more once put on the Web. But numbers of views and downloads are not enough. Effective distance-learning pedagogies that lead the learner through tasks at scales that cannot be achieved in face-to-face classes. A classic challenge for distance learning is "would you teach surgery?" The University of Edinburgh now does just that with support from tutors and assessment, has enabled 1.6 million people (7) to complete university level courses without the need for most initial entry requirements. Teaching at a distance combines media to motivate and entice, including television programs broadcast through the BBC, experiment kits



Support for nontraditional students, team-based quality control, and assessment design are critical.

Laboratory builds a collection of tools to combine remote access, virtual experiments, and citizen science (8) into the curriculum. Advice: Interactions between student-teacher, student-student, and student-materials all can act to support learners (9). Paying attention to the content, and building materials that do the teaching (10), allows direct contact between teacher and learner to be reduced. Structured tasks guide the learner. Working online offers the chance to build in interactivity. Preparation to help learners who need support.

Plan to help learners who need support, and motives; it is not the core. On the other hand, carefully constructed text-based material can feed to the student as if it is speaking to them. Then, using multimedia can build further ways to engage learners in science. Plan to help learners who need support. "Open" is not the same as "free." Openness means accepting those who want to learn as well as those who study to learn. Learning is challenging, and helping students is essential. Some people will manage on their own, but that is not enough for genuinely inclusive education. The self-paced, location-independent properties of online learning make it attractive to the marginalized and those with disabilities (11). Rapid fall-out identified in many MOOCs (12), where only 10% of those who register may complete the course, reflects common challenges. How we approach support for learners influences retention. Early contact with a tutor prevents drop-out, and student attitudes toward the tutor matter (13). For large-scale operation, tutors focus on effective and timely feedback to learners. Support is particularly important as activities start: submission of the first assignment predicts eventual success with a course. Advice: A vital step in coping with access is to recognize the importance of sup-

motivations for participants, how to scale up to genuinely massive access to learning, and how best to assess learning. The opportunity for experimentation gives us the chance to learn more ourselves, as well as to educate others. Distance universities, from their inception, ran courses for thousands of learners, accepted open entry, and led the move into online methods of teaching and learning.

**References and Notes**  
1. J. B. Aronson, *Science*, 2012, 336, 1200-1201.  
2. C. A. Lounsbury, *J. Educ. Res.*, 2012, 114, 100-101.  
3. S. J. Liebowitz, *J. Econ. Surv.*, 2012, 26, 1-10.  
4. S. J. Liebowitz, *J. Econ. Surv.*, 2012, 26, 1-10.  
5. S. J. Liebowitz, *J. Econ. Surv.*, 2012, 26, 1-10.  
6. S. J. Liebowitz, *J. Econ. Surv.*, 2012, 26, 1-10.  
7. S. J. Liebowitz, *J. Econ. Surv.*, 2012, 26, 1-10.  
8. S. J. Liebowitz, *J. Econ. Surv.*, 2012, 26, 1-10.  
9. S. J. Liebowitz, *J. Econ. Surv.*, 2012, 26, 1-10.  
10. S. J. Liebowitz, *J. Econ. Surv.*, 2012, 26, 1-10.  
11. S. J. Liebowitz, *J. Econ. Surv.*, 2012, 26, 1-10.  
12. S. J. Liebowitz, *J. Econ. Surv.*, 2012, 26, 1-10.  
13. S. J. Liebowitz, *J. Econ. Surv.*, 2012, 26, 1-10.

We regret to inform you that your paper has been rejected

The problem is not well studied.

A first step to understand statement strength is to distinguish strong and weak statements.

Statement strength is inherently relative.



# Authors post latex source for different versions of the same paper

arXiv.org > math > arXiv:1109.4363

Mathematics > Probability

## The Segregated Lambda-coalescent

Nic Freeman

*(Submitted on 20 Sep 2011 (v1), last revised 2 Nov 2013 (this version, v3))*

### Submission history

From: Nic Freeman [[view email](#)]

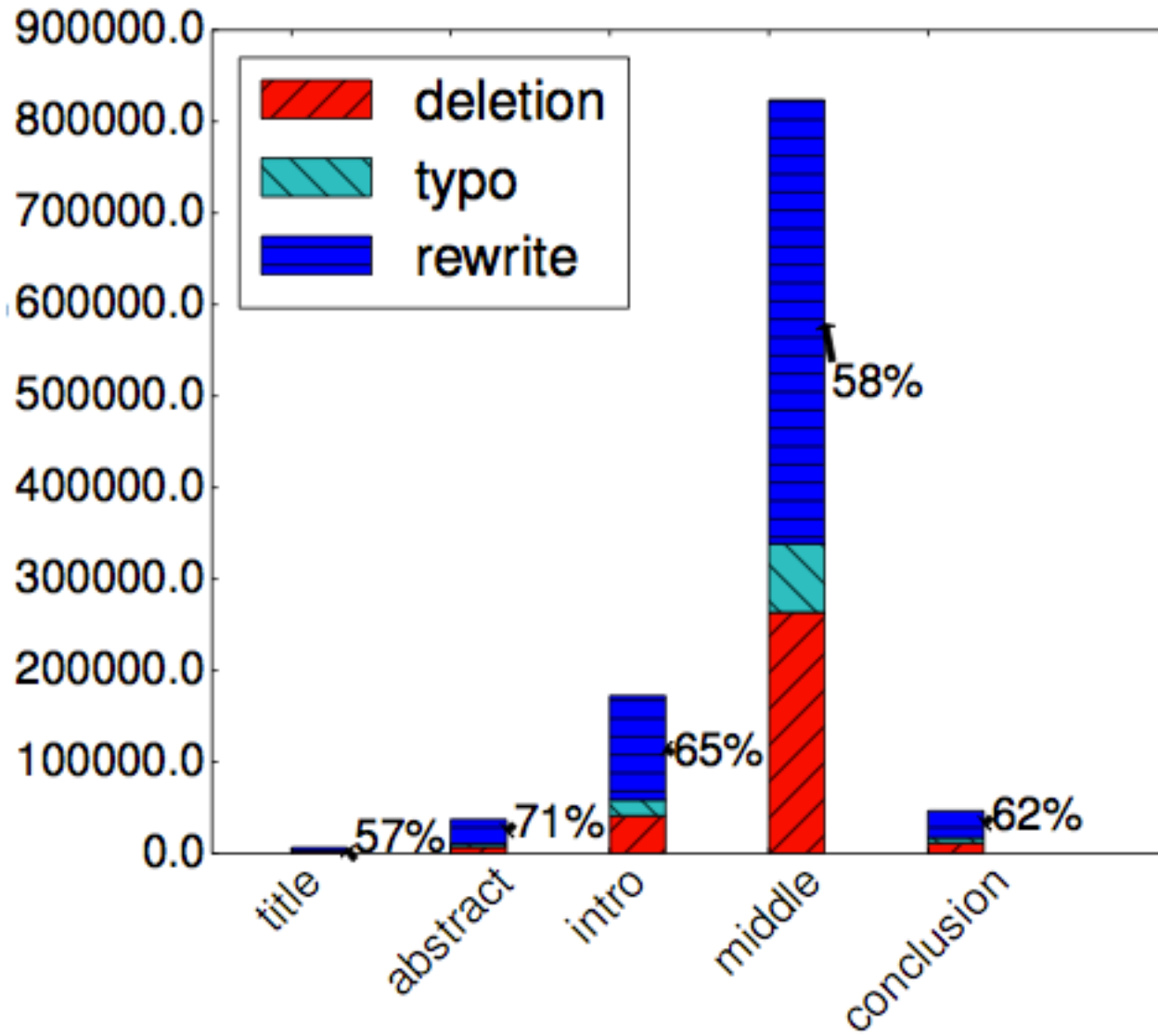
[v1] Tue, 20 Sep 2011 17:13:33 GMT (125kb,D)

[v2] Wed, 9 Nov 2011 15:03:40 GMT (137kb,D)

[v3] Sat, 2 Nov 2013 22:15:38 GMT (108kb,D)

Is it only typos?

# A lot of rewrites are made between different versions



# Align different versions of the same paper to find sentence pairs

[Barzilay and Elhadad 2003]

## Phase transitions in a spatial coalescent

Nic Freeman

(Submitted on 20 Sep 2011 (this version), latest version 2 Nov 2013 (v3))

We construct a natural extension of the Lambda-coalescent to a spatial continuum, and analyse its behaviour.

Like the Lambda-coalescent, at any time  $t > 0$  the individuals in our model can be separated into (i) a dust component and (ii) large blocks of coalesced individuals. We identify a five phase system, where our phases are defined according to changes in the qualitative behaviour of the dust and blocks. We completely classify the phase behaviour, and obtain necessary and sufficient conditions for the model to come down from infinity.

## The Segregated Lambda-coalescent

Nic Freeman

(Submitted on 20 Sep 2011 (v1), last revised 2 Nov 2013 (this version, v3))

We construct an extension of the Lambda-coalescent to a spatial continuum and analyse its behaviour. Like the Lambda-coalescent, the individuals in our model can be separated into (i) a dust component and (ii) large blocks of coalesced individuals. We identify a five phase system, where our phases are defined according to changes in the qualitative behaviour of the dust and large blocks. We completely classify the phase behaviour, including necessary and sufficient conditions for the model to come down from infinity.

# Examples of potential strength changes

The algorithm is *studied* in this paper .

... circadian pattern and burstiness in *human communication activity* .

The algorithm is *proposed* in this paper .

... circadian pattern and burstiness in *mobile phone communication* .

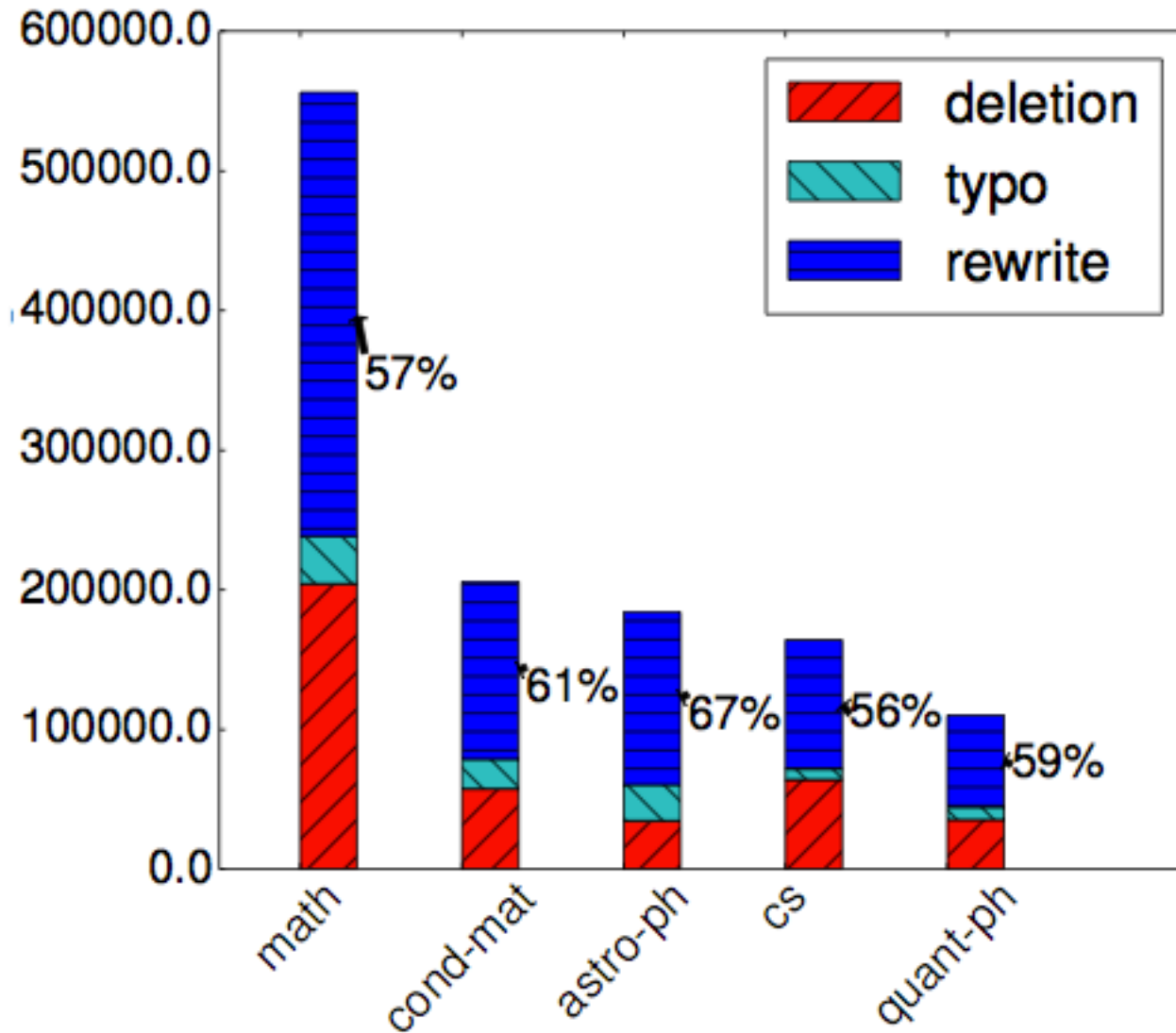
# Examples of potential strength changes

they maximize the expected revenue of the seller but *induce efficiency loss* .

they maximize the expected revenue of the seller but *are inefficient* .



# Top categories in making changes



## A corpus of sentence-level revisions focusing on potential strength changes

- 108K pairs from abstracts or introductions
  - similarity score for the pair was larger than 0.5
- Final labeling instructions:  
stronger, weaker, no strength change, I can't tell
- Labeled 500 pairs on Amazon Mechanical Turk
  - 9 labels and *COMMENTS* each

# Overall labeling results

- Among the 500 pairs, Fleiss' Kappa was 0.242, which indicates fair agreement
- 386 pairs have an absolute-majority label  
Fleiss' Kappa is 0.322, and 74.4% of pairs were strength changes  
(93 weaker, 194 stronger, 99 no change)
- Most labels agree with our intuitions, but there are also some differences

# Participants are swayed by specificity

S1: ... using data from numerics and experiments .

S2: ... using data sets from numerics in the point particle limit and one experimental data set .

S2 is stronger: “S2 is more specific in its description which seems stronger.”

S2 is weaker: “‘one experimental data set’ weakens the sentence”

**Similar findings in courts** [Bell and Loftus (1989)]

# Participants interpret constraints/conditions not in strictly logical ways

S1: we also proved that if  
[MATH] is sufficiently  
homogeneous then ...

S2: we also proved that if  
[MATH] is *not totally*  
*disconnected* and sufficiently  
homogeneous then ...

(stronger) We have more detail/proof in S2

(stronger) the words "not totally disconnected" made the sentence sound more impressive.

# Participants can have a different understanding of domain-specific terms

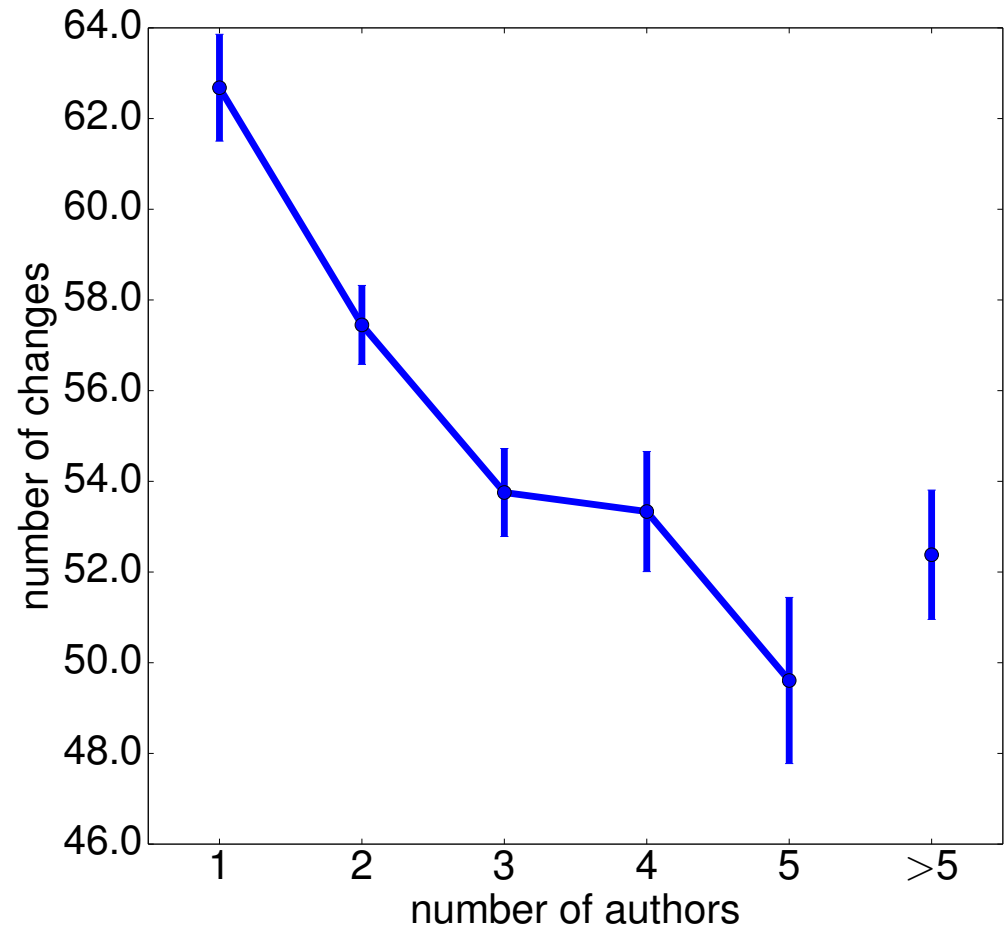
S1: in the current paper we *discover* several variants of qd algorithms *for* quasiseparable matrices .

S2: in the current paper we *adapt* several variants of qd algorithms *to* quasiseparable matrices .

S2 is stronger: “in S2 Adapt is stronger than just the word discover. adapt implies more of a proactive measure.”

# This type of corpus can enable other interesting studies

The more authors,  
the fewer changes!



- The labels and *comments* we collected can hopefully provide insights into better ways to define and approach this problem.
- The ultimate goal of this study is to understand the effects of statement strength on the public, which can lead to various applications in public communication.



We confirm that language matters via natural experiments, and show that this paradigm can also improve prediction performance

We collect the first large-scale dataset on language strength

Twitter Data

<http://chenhaot.com/pages/wording-for-propagation.html>

Twitter Demo

<http://chenhaot.com/retweetedmore>

Twitter Quiz

<http://chenhaot.com/retweetedmore/quiz>

<http://www.nytimes.com/interactive/2014/07/01/upshot/twitter-quiz.html>

Strength data

<http://chenhaot.com/pages/statement-strength.html>

**I hope this is the beginning of an interesting journey!**