Inferring Interestingness in Online Social Networks

Will M. Webberley
Stuart M. Allen
Roger M. Whitaker

School of Computer Science & Informatics

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

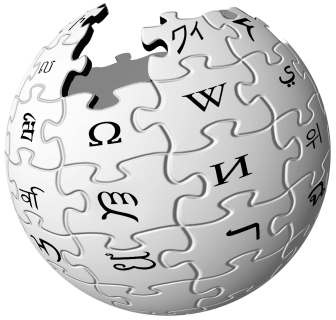
Information retrieval

- Used by everyone every day for varied purposes

Google

YouTube

bing



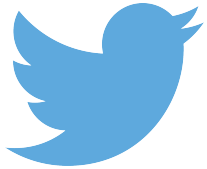
BBC NEWS

Online social networks

- *Also* used by everyone every day for varied purposes
- Information 'quick-fix'





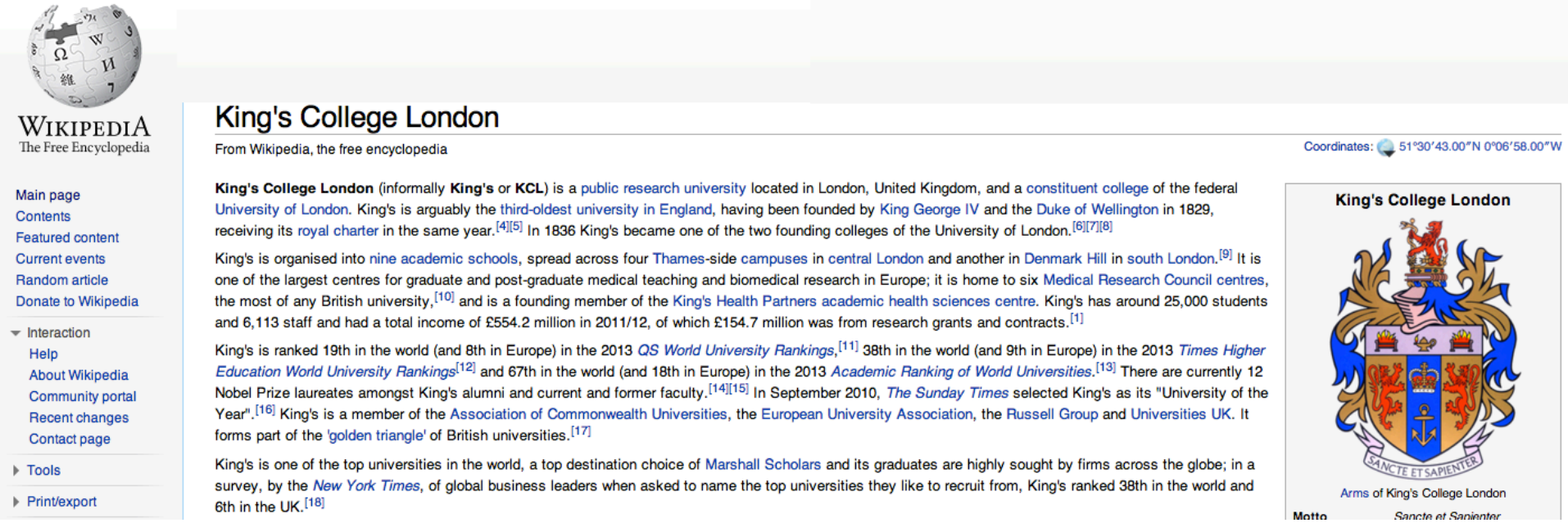




Online social network services

Information-retrieval system services

Epistemic search

search information just to satisfy desire for knowledge[1]



WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools

Print/export

King's College London

From Wikipedia, the free encyclopedia

Coordinates: 51°30′43.00″N 0°06′58.00″W

King's College London (informally **King's** or **KCL**) is a public research university located in London, United Kingdom, and a constituent college of the federal University of London. King's is arguably the third-oldest university in England, having been founded by King George IV and the Duke of Wellington in 1829, receiving its royal charter in the same year.[4][5] In 1836 King's became one of the two founding colleges of the University of London.[6][7][8]

King's is organised into nine academic schools, spread across four Thames-side campuses in central London and another in Denmark Hill in south London.[9] It is one of the largest centres for graduate and post-graduate medical teaching and biomedical research in Europe; it is home to six Medical Research Council centres, the most of any British university,[10] and is a founding member of the King's Health Partners academic health sciences centre. King's has around 25,000 students and 6,113 staff and had a total income of £554.2 million in 2011/12, of which £154.7 million was from research grants and contracts.[1]

King's is ranked 19th in the world (and 8th in Europe) in the 2013 *QS World University Rankings*,[11] 38th in the world (and 9th in Europe) in the 2013 *Times Higher Education World University Rankings*[12] and 67th in the world (and 18th in Europe) in the 2013 *Academic Ranking of World Universities*.[13] There are currently 12 Nobel Prize laureates amongst King's alumni and current and former faculty.[14][15] In September 2010, *The Sunday Times* selected King's as its "University of the Year".[16] King's is a member of the Association of Commonwealth Universities, the European University Association, the Russell Group and Universities UK. It forms part of the 'golden triangle' of British universities.[17]

King's is one of the top universities in the world, a top destination choice of Marshall Scholars and its graduates are highly sought by firms across the globe; in a survey, by the *New York Times*, of global business leaders when asked to name the top universities they like to recruit from, King's ranked 38th in the world and 6th in the UK.[18]

King's College London

SANCTE ET SAPIENTER

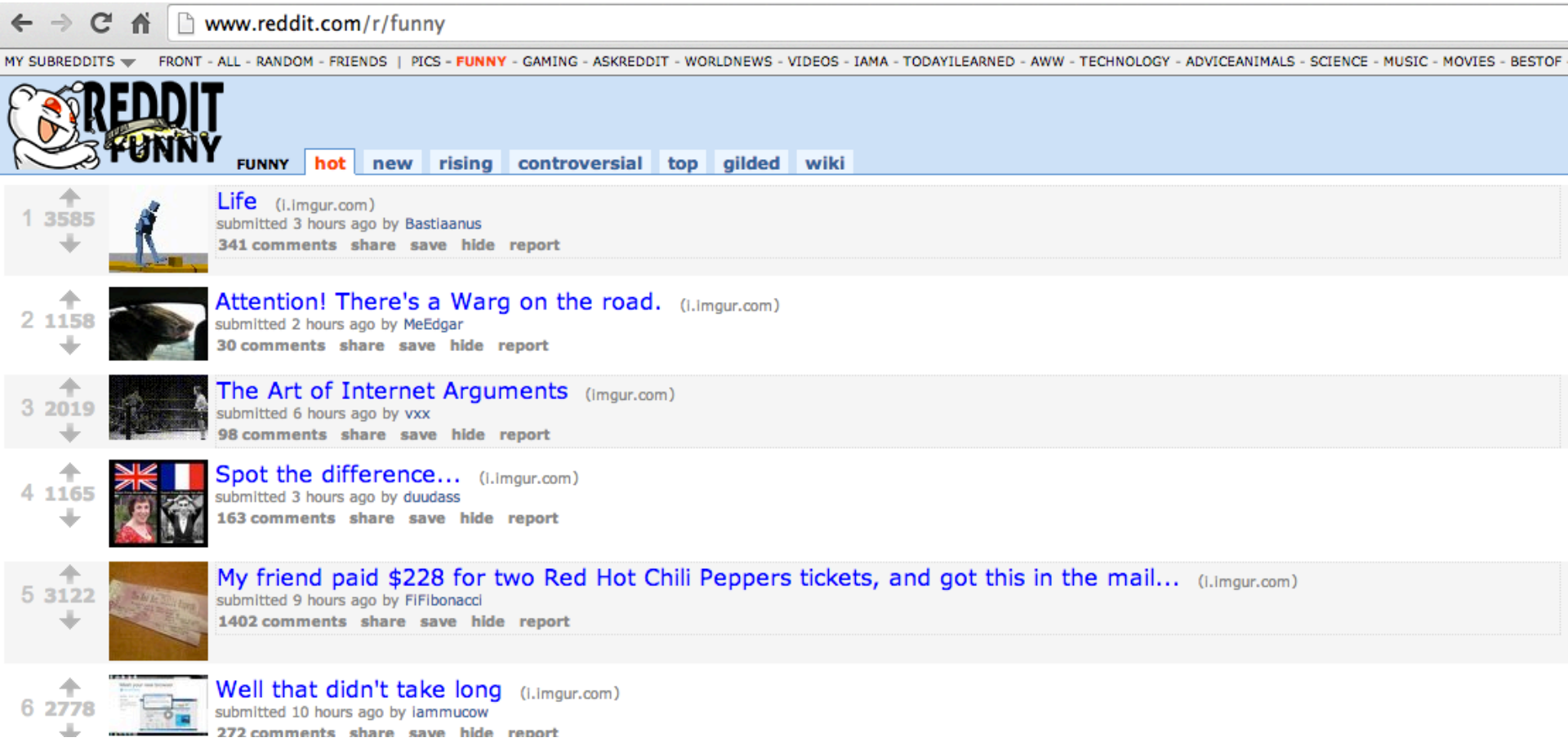
Arms of King's College London

Motto *Sancte et Sapienter*

[1] Yunjie Xu, Relevance Judgment in Epistemic and Hedonic Information Searches, Journal of the American Society for Information Science and Technology, 2007

Hedonic search

search information for fun or affective stimulation[1]



Effective stimulation

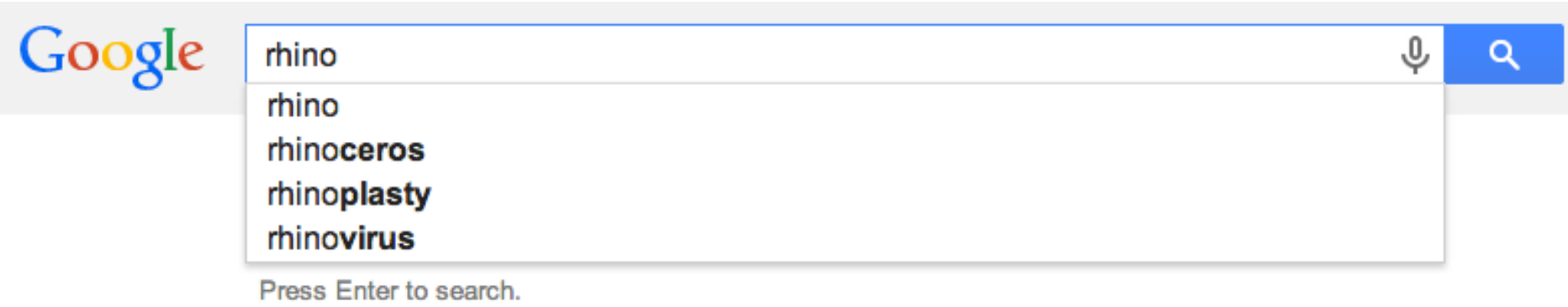
what draws you in?

interest relevance emotion thought-inducing provocative

...

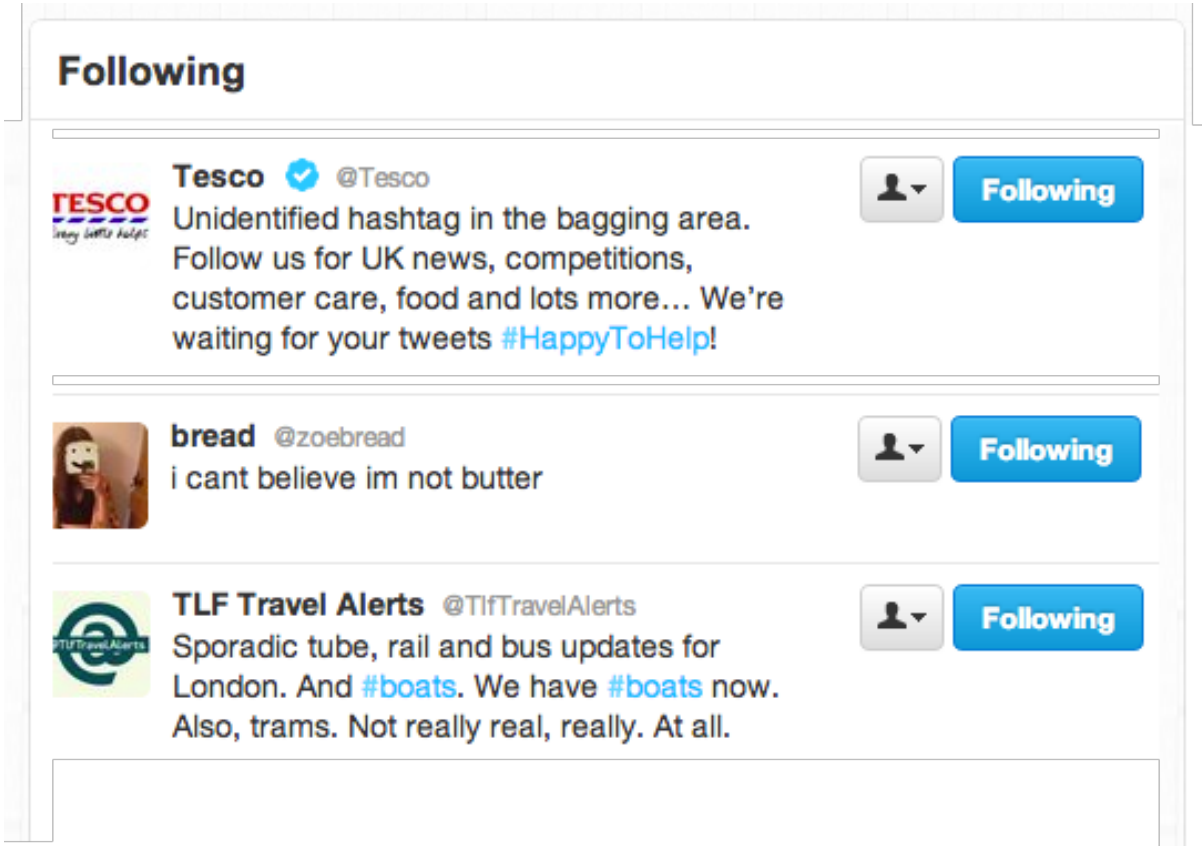
“Retrieving” from information-retrieval systems

- Usually to solve an immediate ‘problem’
- Provide a search term



“Retrieving” from OSNs

- Typically hedonic
- Not to solve an immediate problem
- Provide a set of users to receive information from
- Don’t know what you will receive



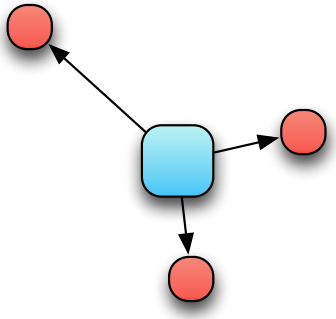
‘Noisy’ information

- Non-relevant information from information-retrieval systems
- Non-interesting information from online social networks

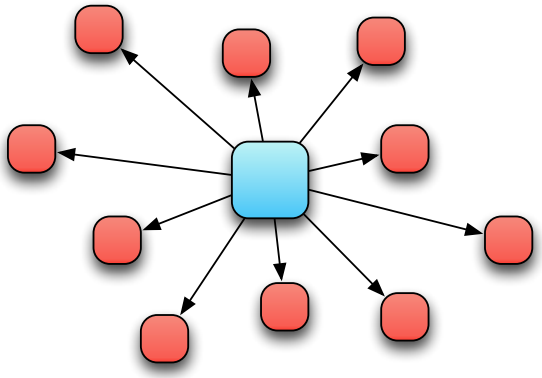
not everyone you follow will always produce interesting content

- Effectively stimulating information is ‘interesting’

Some level of control…



noise recall precision



noise recall precision

How to identify interesting information?

we want affective stimulation

Can we identify and deliver **interesting** and **relevant** information to users…

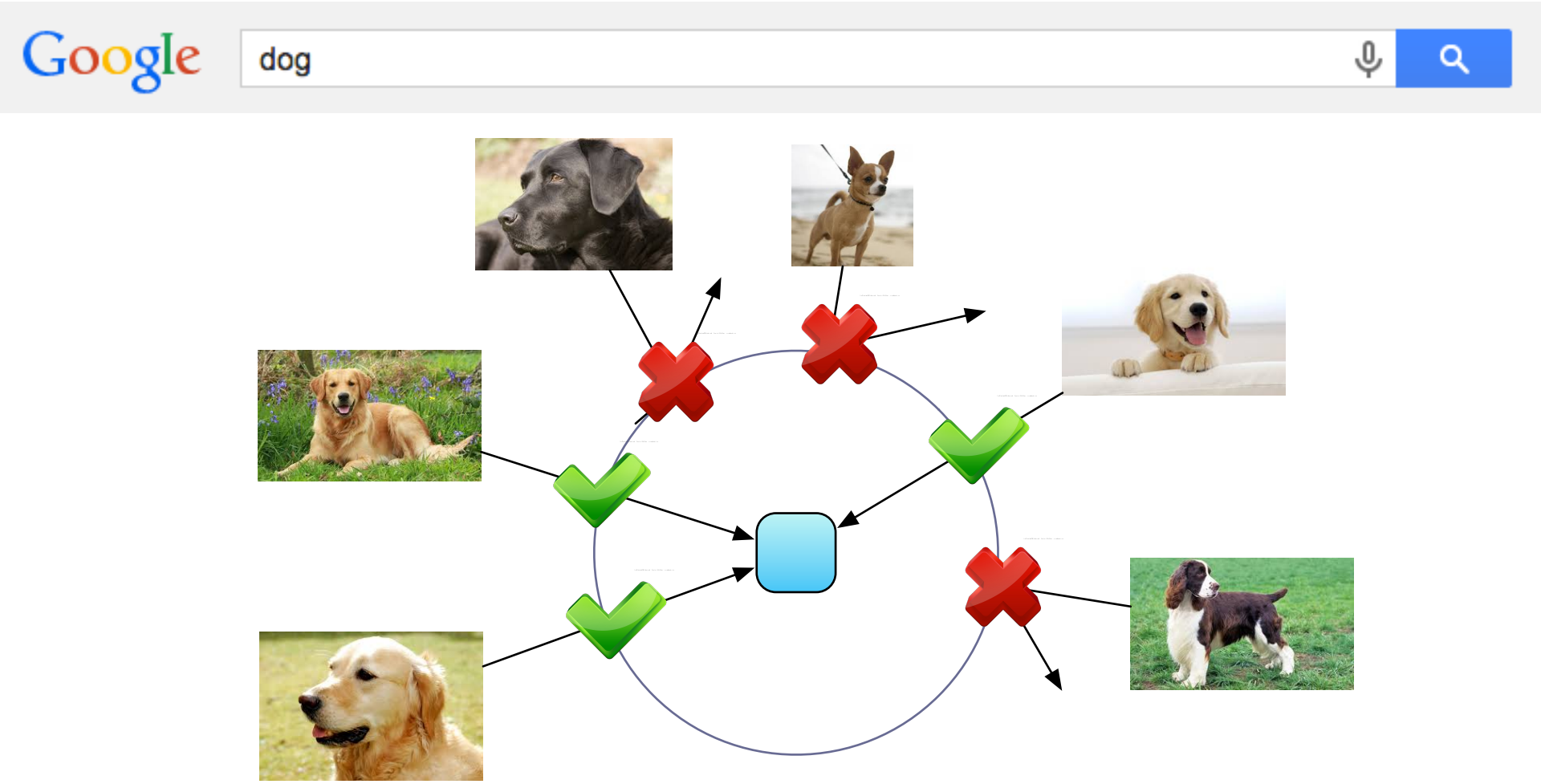
… yet without them having to **look** for it or even **know** about it first?

Can we identify and deliver **interesting** and **relevant** information to users…

… yet without them having to **look** for it or even **know** about it first?

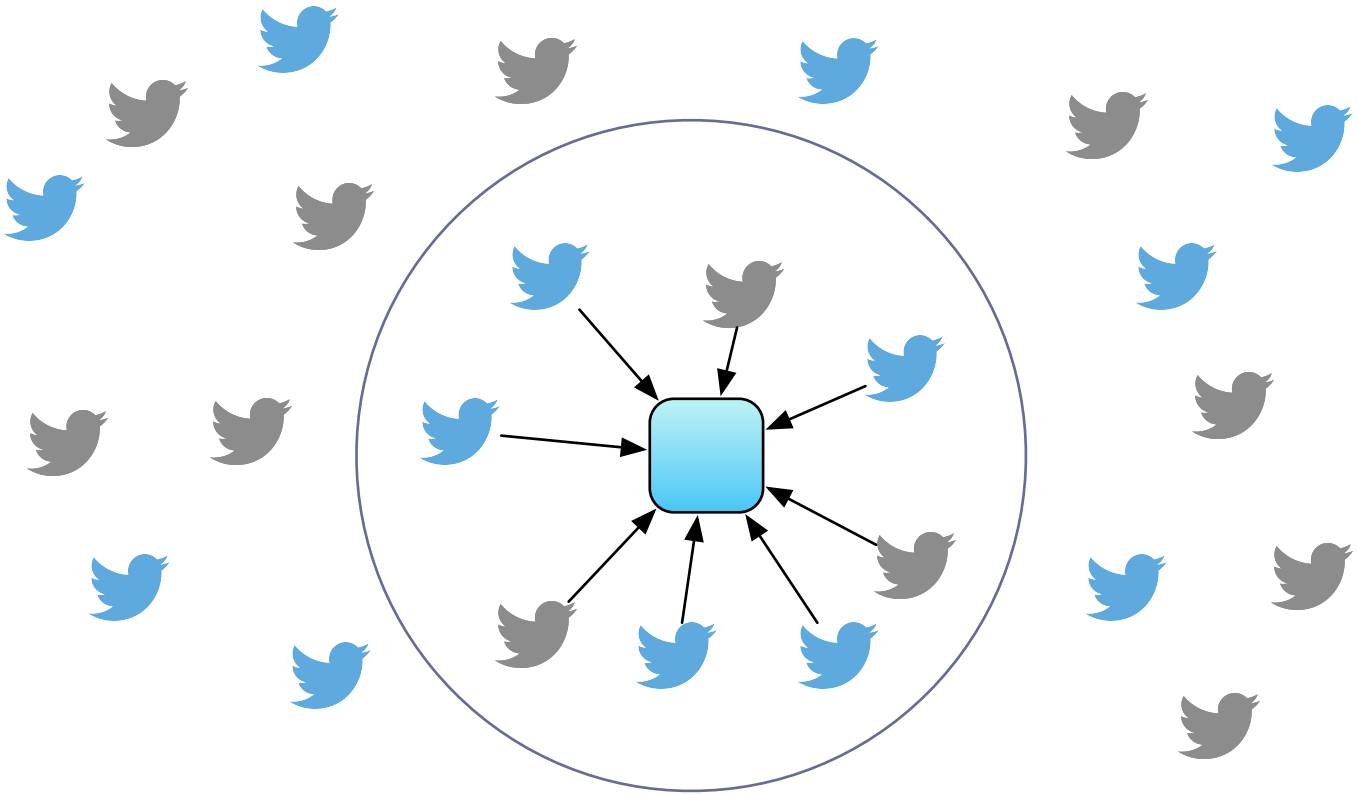
Google's filter 'bubble'

what are we not getting to see?



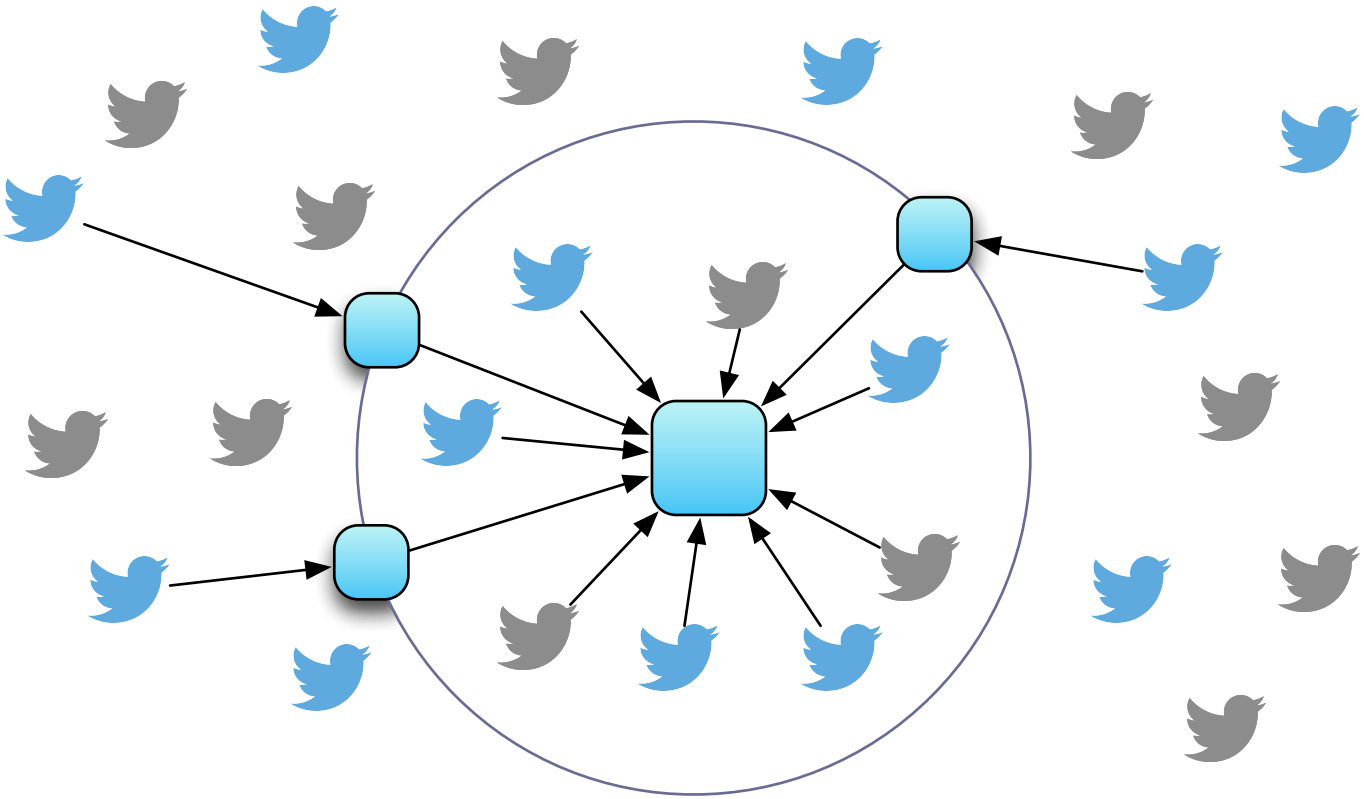
Similarly…

what are we not getting to see or know about?



Retweets

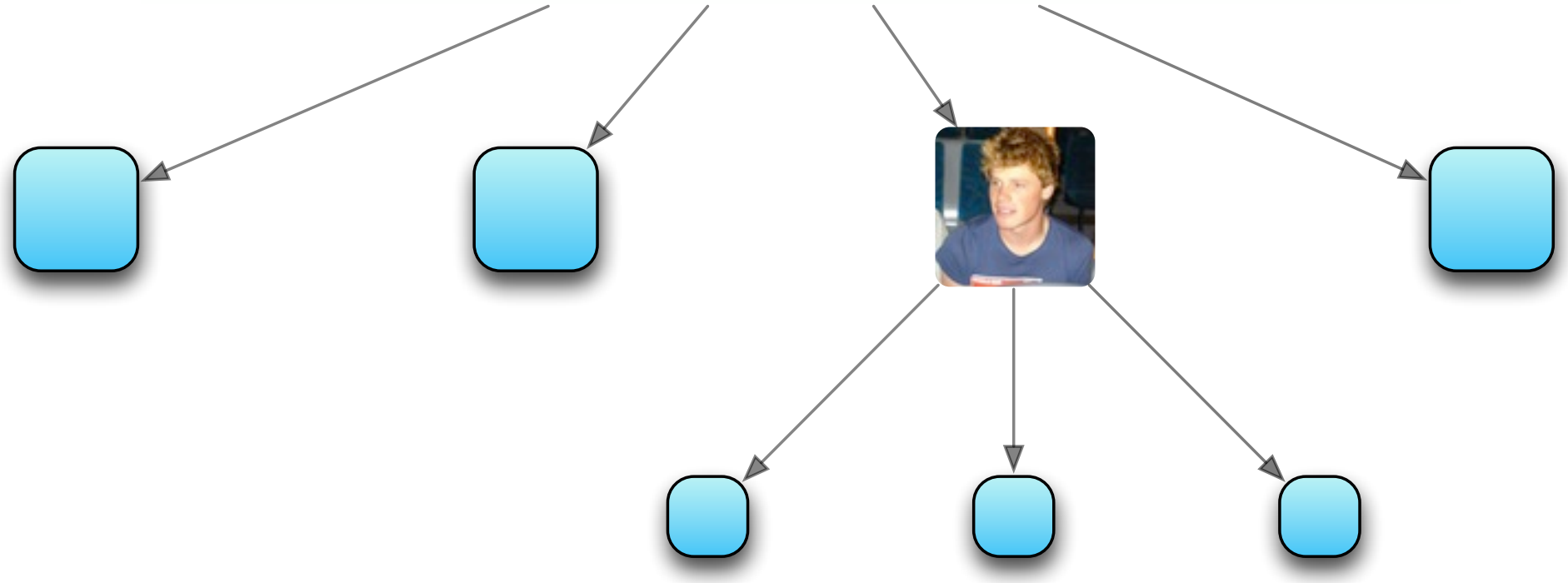
what do we have the potential to see?



Retweets

Karissa McKelvey @karissamck

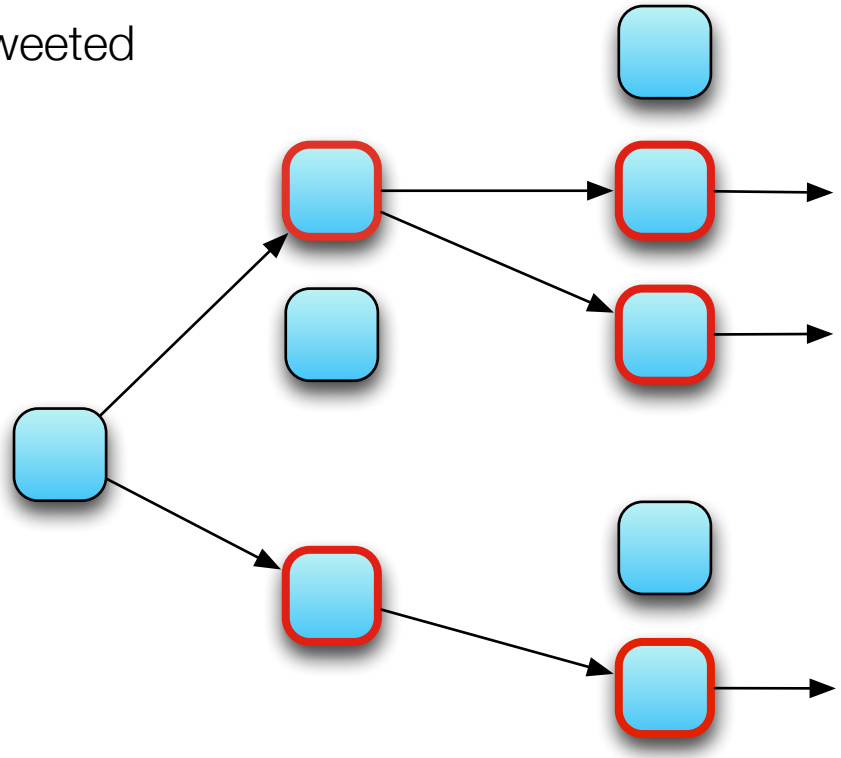
"We didn't read half the papers we cite because they are behind a paywall." #ACM #overlyhonestmethods



Retweets

- Decentralised information propagation
- Retweeted Tweets tend to have a higher quality (voting / endorsement)
- Can, themselves, also be retweeted
- Friends are information filters

Retweet count = 5



Retweet count

retweet count = raw popularity

retweet count ≠ interestingness

Retweet count

retweet count ≠ interestingness

BBC News (World) @BBCWorld

"Sweetie I'm coming - I'm gonna get you out of here" - man rescues children from school flattened by tornado.

56	23
RETWEETS	FAVORITES

Justin Bieber @justinbieber

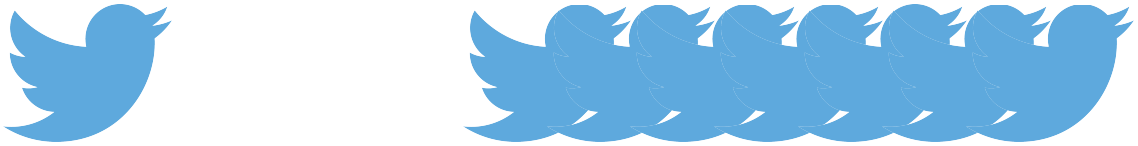
all love

108,485	64,503
RETWEETS	FAVORITES

Initial research

- Further understanding of propagation in Twitter
- Retweet properties and behaviours
- Can retweets be used as a basis for estimating interestingness?
- Research based on a set of ~26,000 Tweets from the public timeline

Retweet ‘groups’



$$RG(t) = t + RT(t)$$

Properties include

Path-length (propagation *depth*)

Size ($|RG(t)|$)

Audience (set of recipient users of an instance of t)

n degrees of separation

‘Real’ world

6 [2]

OSN (Facebook)

~4.7 [3]

[2] ‘Real’ world experiment by Stanley Milgram, 1967

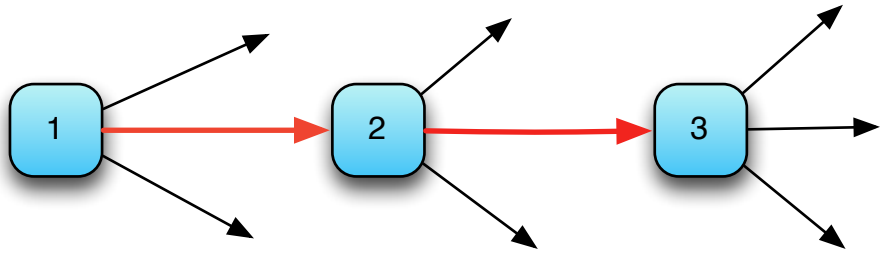
[3] Analysis of Facebook by Backstrom et al., 2011

Path-length

penetration of propagation

A (re)tweet by user @user3 originally authored by @user1:

```
RT @user2: RT @user1: here is the Tweet content!
```

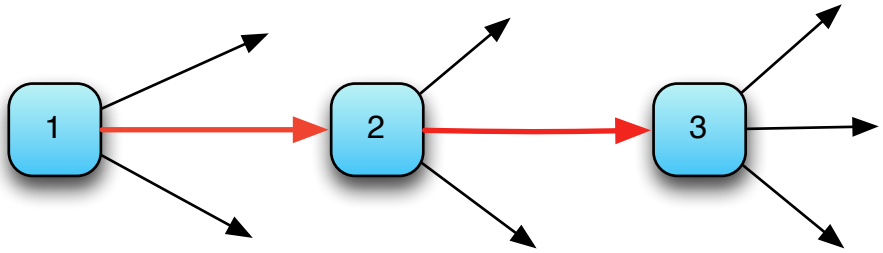


Has path-length of 2

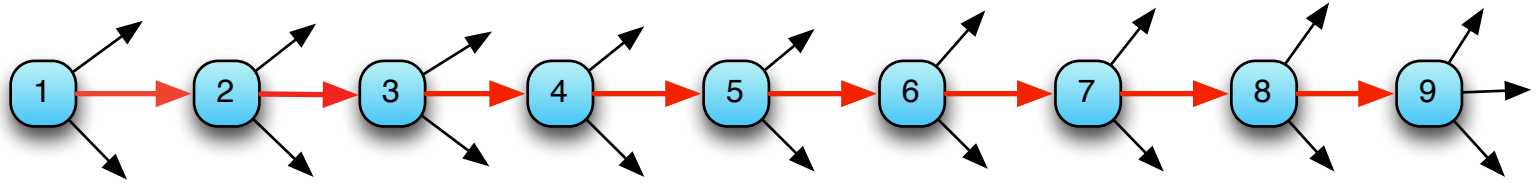
Path-length

penetration of propagation

Average observed path-length: 2



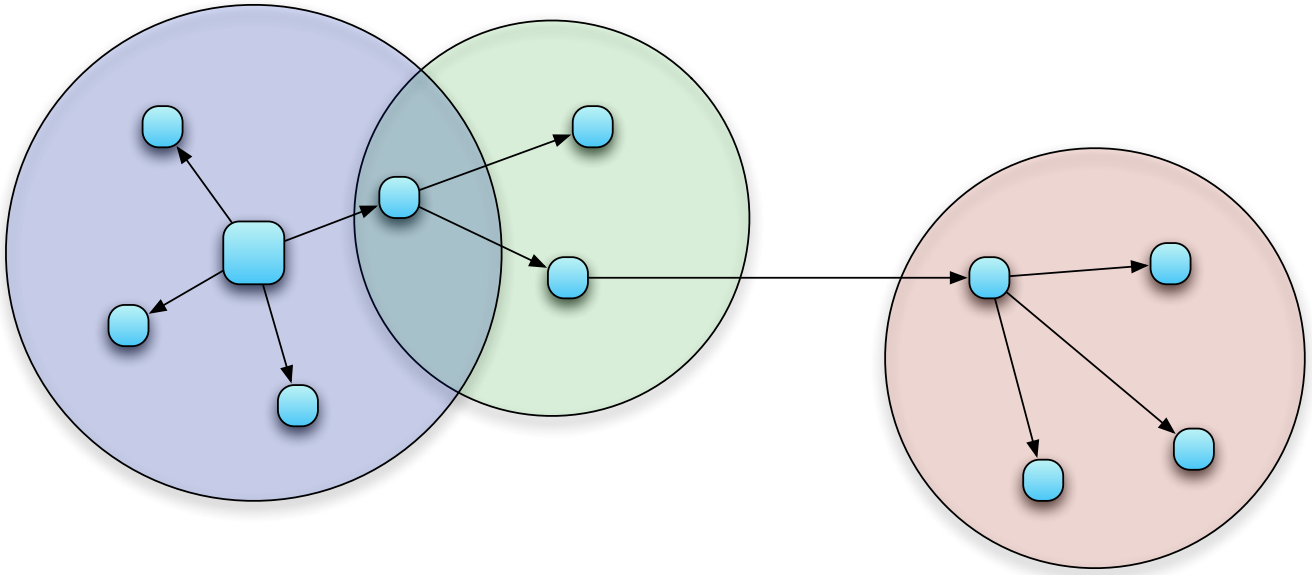
Longest observed path-length: 9



Audience

Tends to increase with increases in path-length

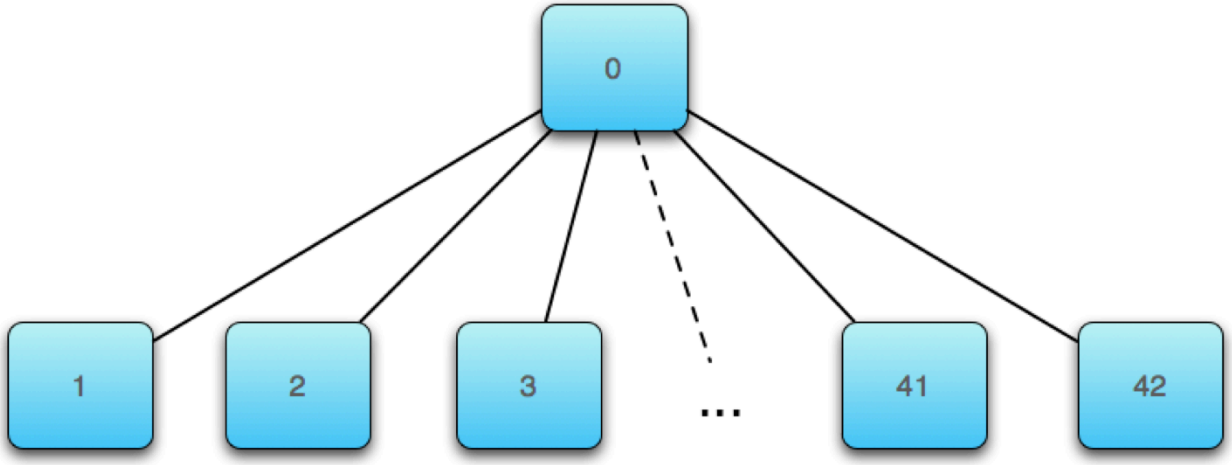
However, the social graph exhibits follower overlap within communities



Audience overhead

who sees the Tweet more than once?

Strong relationships with the social graph

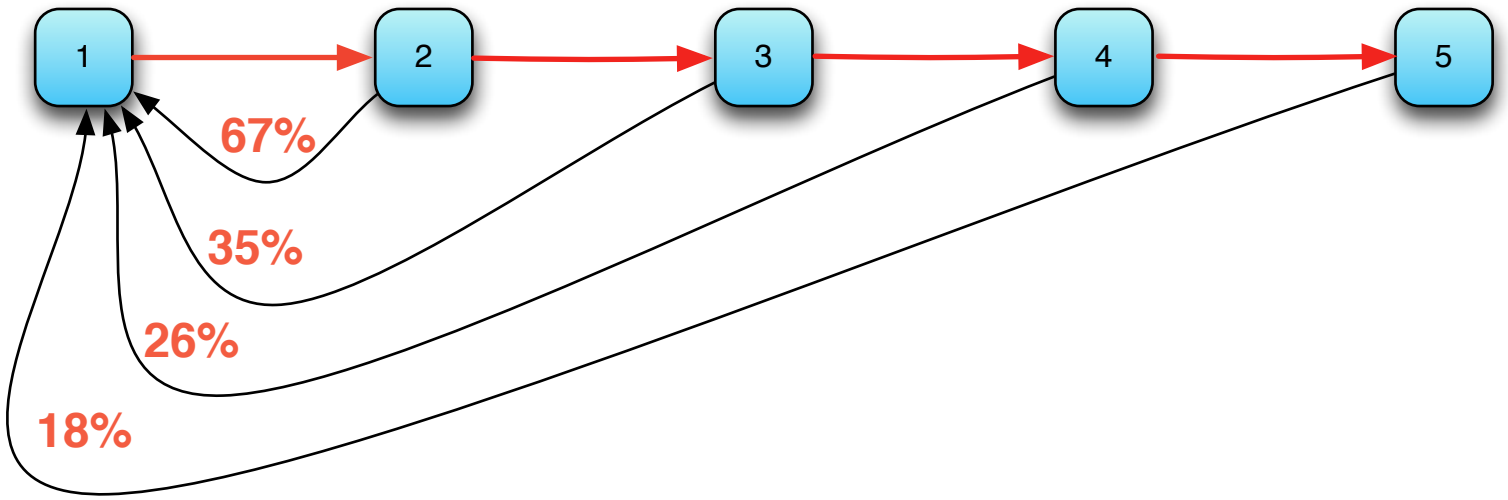


Audience = 7,327

Overhead = 48,868 (667%)

More relationships…

… indicate retweet chains ‘mirror’ aspects of the social graph



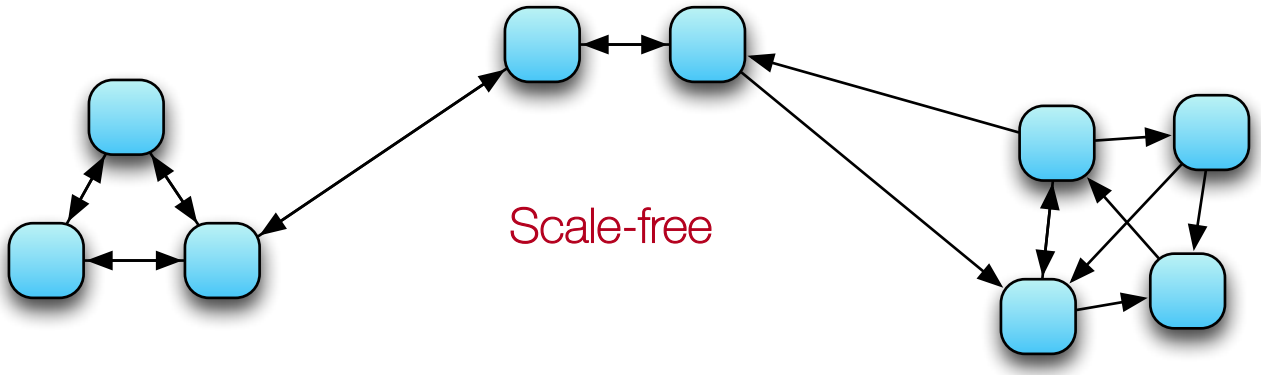
Importance of social structure

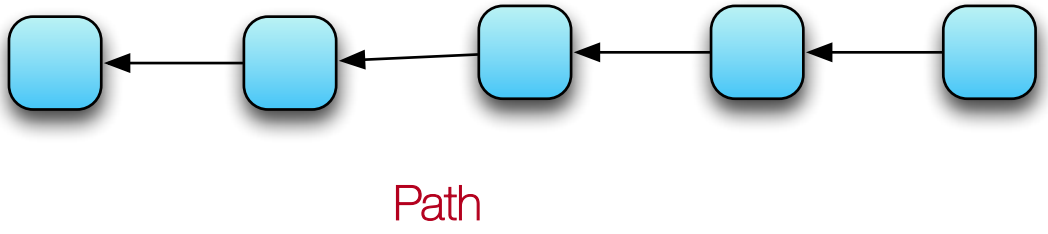
Clearly important in Tweet propagation

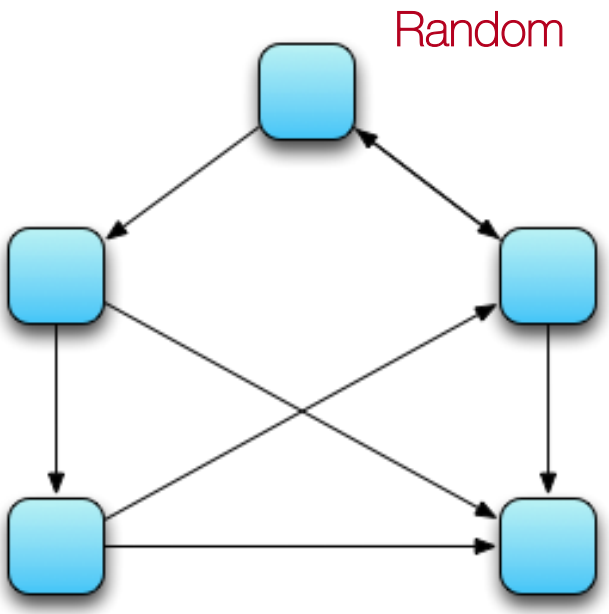
Can it be used to predict Tweet propagation?

Can predictions be used to estimate Tweet interestingness?

Social structure analysis



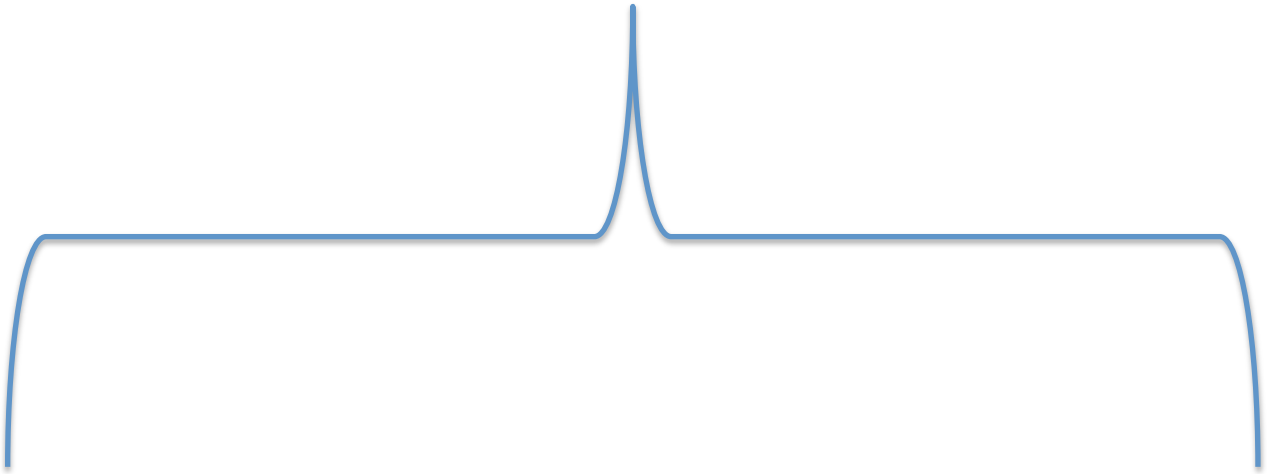




Retweet *decisions*

- Implies a particular user's interest in the Tweet
- Culmination of many decisions -> Tweet becomes *more* interesting

Retweet count



Interestingness

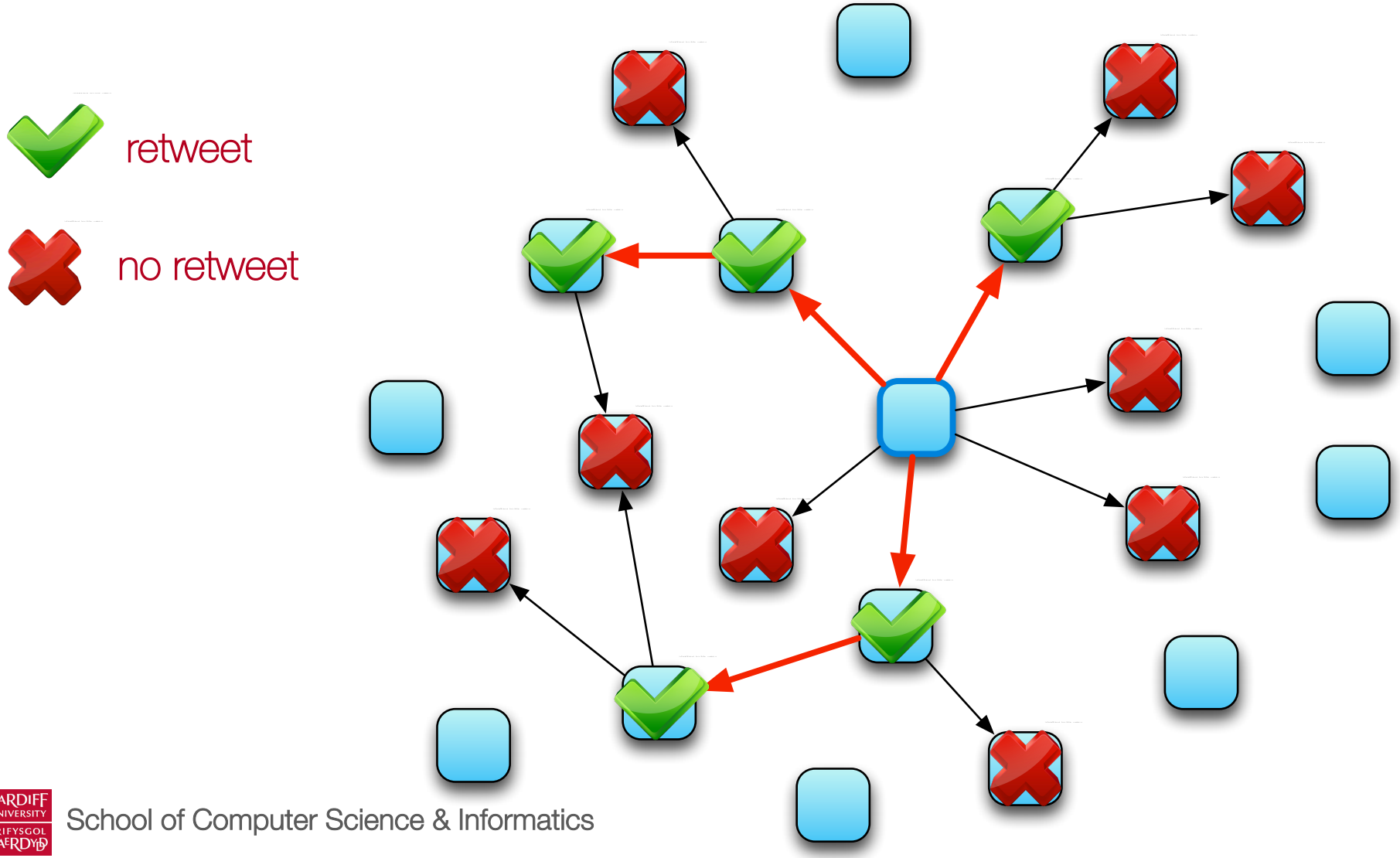
Retweet decisions

User influence

Follower count,
Tweet frequency,
etc.

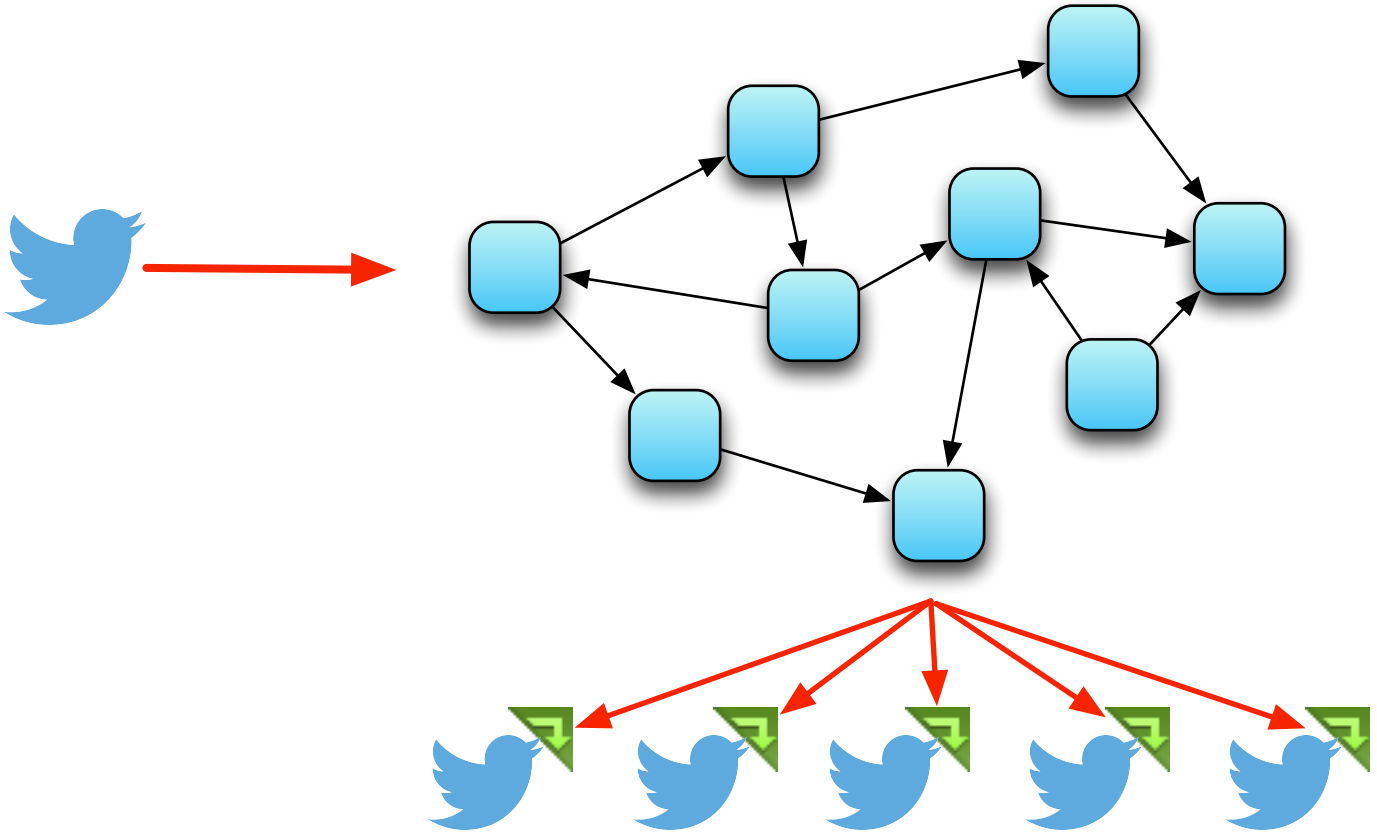
Social structure analysis

simulating retweet decisions



Social structure analysis

how many retweets are produced by the graph?



Model retweet *decisions*

Collect ~12,000 Tweets

Extract features (**`has_url`**, **`is_mentioned`**, **`retweeter_follows`**, etc.)

Train logistic regression and introduce a Tweet, ***t***

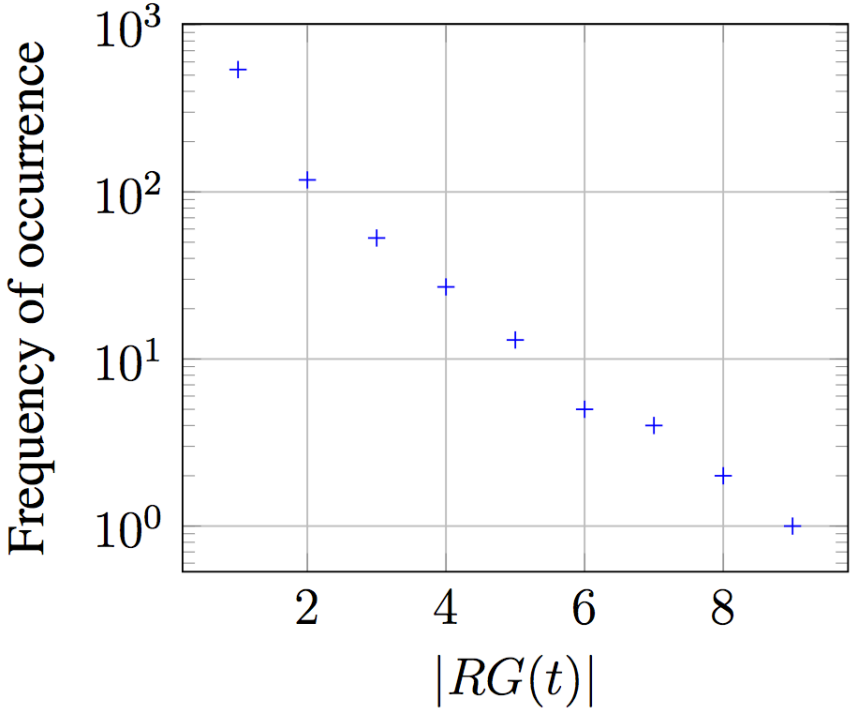
Predict probability of a receiver's retweet decision on ***t***

Culmination of decisions provides a retweet count for ***t***

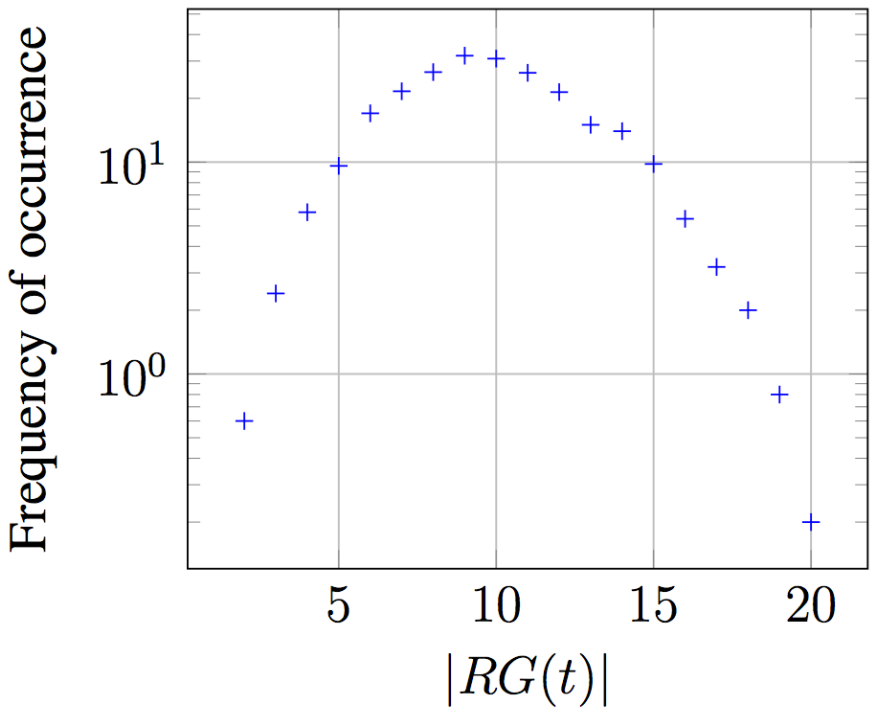
Social graph structure analyses

retweet count distributions

Path network

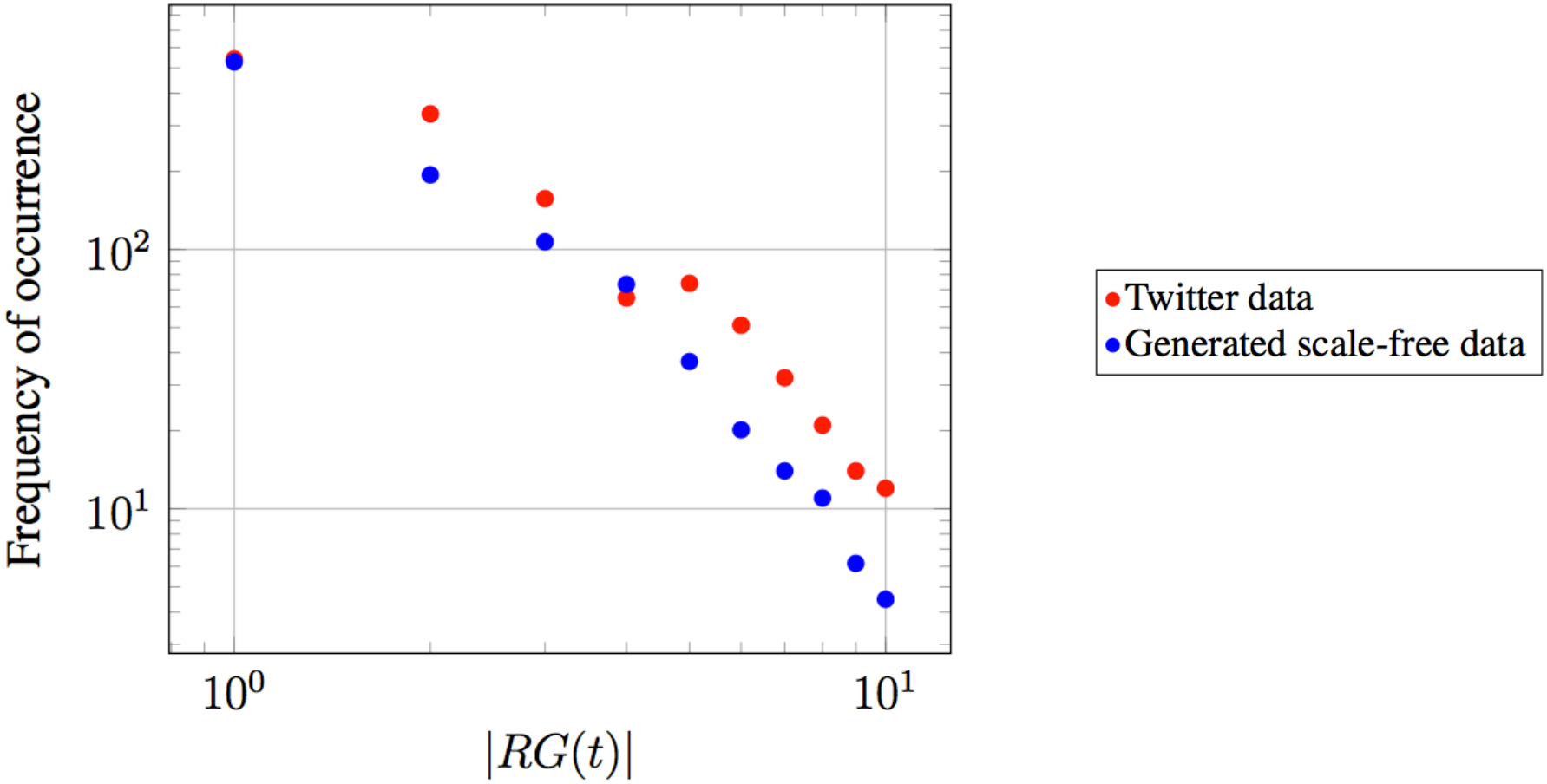


Random network



Social graph structure analyses

retweet count distributions



Social graph structure analyses

Clearly a significant impact

Can it be used as a basis for estimating interestingness?

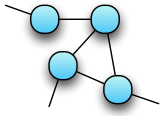
Inferring interestingness

Comparing an expected retweet count to the observed retweet count

A hypothesis

If observed greater than expected, then there's something that makes the Tweet particularly interesting

Observed count



Expected count

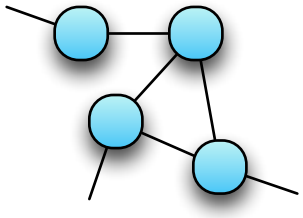
Inferring interestingness

	Tweet 1	Tweet 2
author	Cardiff University (@cardiffuni)	Cardiff University (@cardiffuni)
contains hashtag	✔	✔
contains URL	✔	✔
contains smiley	✗	✗
Retweet count:	238	11

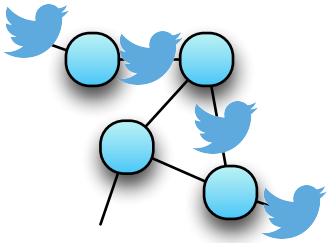
Inferring interestingness

producing the estimated retweet count

Collect a user's local network and Tweets



Build a copy of the network locally and introduce Tweets



Simulate Tweets' retweet decisions within the network

Inferring interestingness

Crawled Twitter for data and ran simulations for each user's Tweets

Resulted in expected retweet counts for 10,000 Tweets

Each Tweet can now be labelled as interesting or non-interesting

… but does it work?

Initial verifications of the labels

Crowdsourced through Amazon's Mechanical Turk

Asked: *"Of these 5 Tweets, which is the most interesting?"*

Low accuracy: 33% agreement on positively-labelled Tweets

Other issues

Inefficient:

Data collection (API calls)

Takes a lot of time

Limited to a small subset of users and Tweets

Inaccurate anyway!

What next?

Idea is there, but need:

To find a better way for estimating the expected retweet count

To apply mechanism to a wider range of Tweets

More efficiency (time, data collection, etc.)

To consider Tweet *ranking*

A more direct solution

Train a model to directly predict an expected retweet count

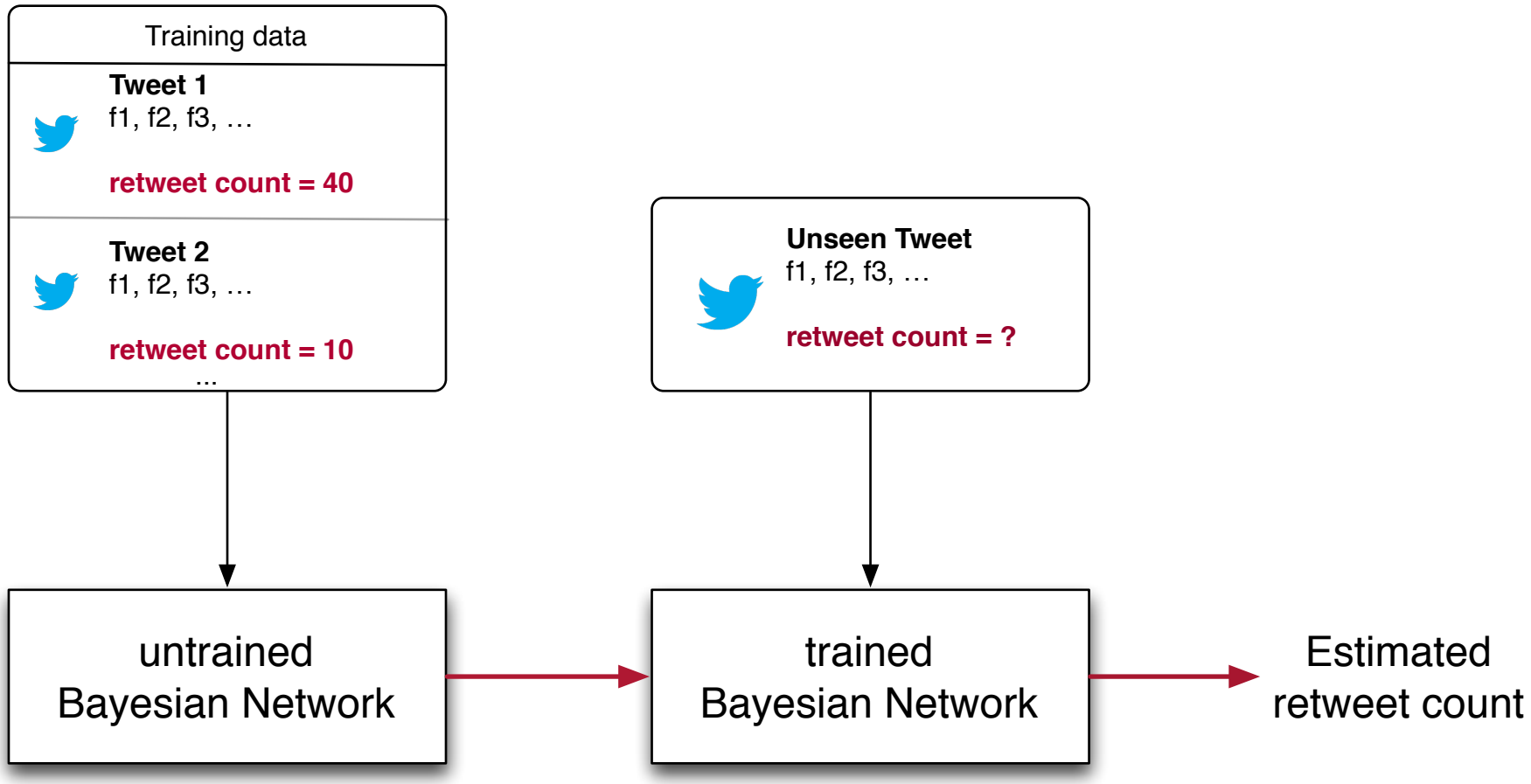
No need for collection and simulation of a local network…

… but will need a more substantial set of features

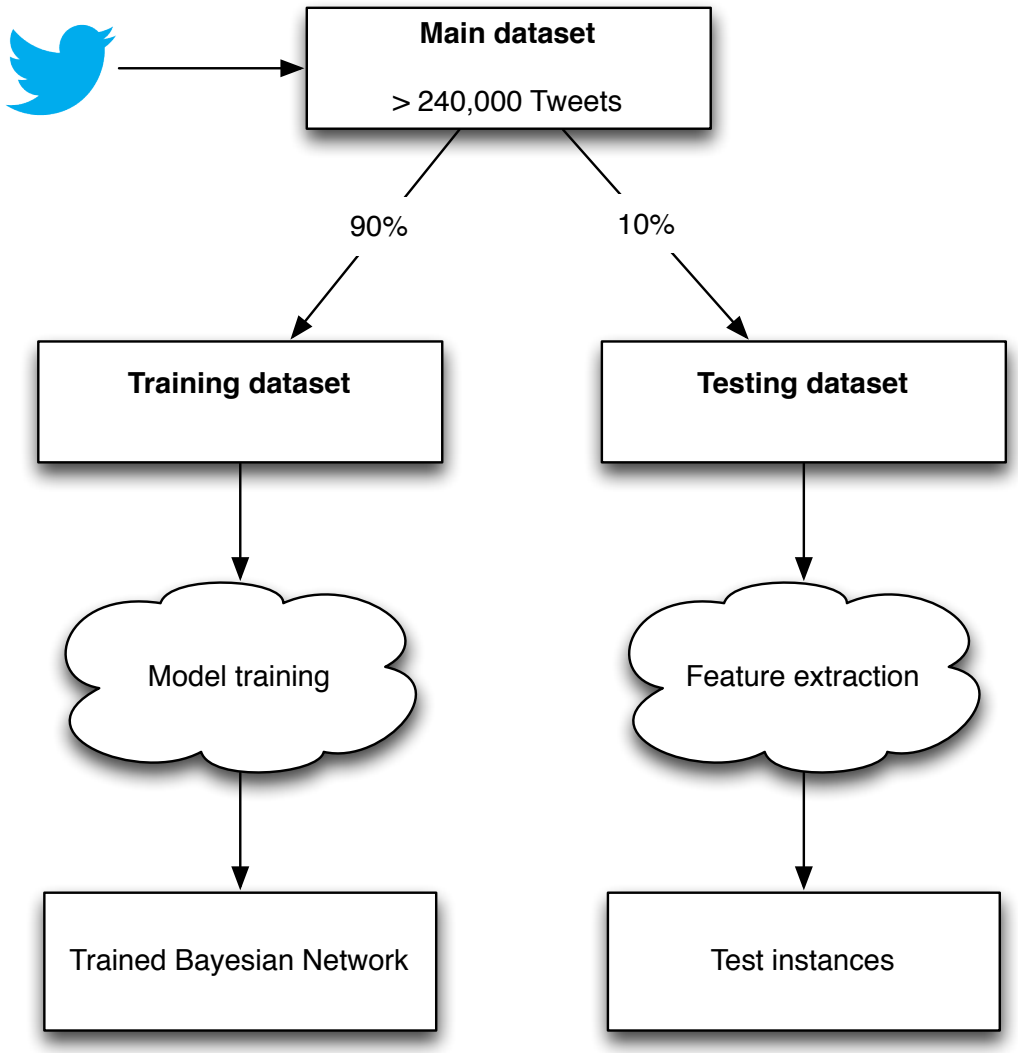
Features

genetics	‘Tweet-etics’
genome	Tweet features
eye colour, height, personality, etc.	text length, contains URL, hashtags, etc.
environment	network features
temperature, humidity, peers, etc.	author followers, community interest, etc.
	Total = 31

Action plan

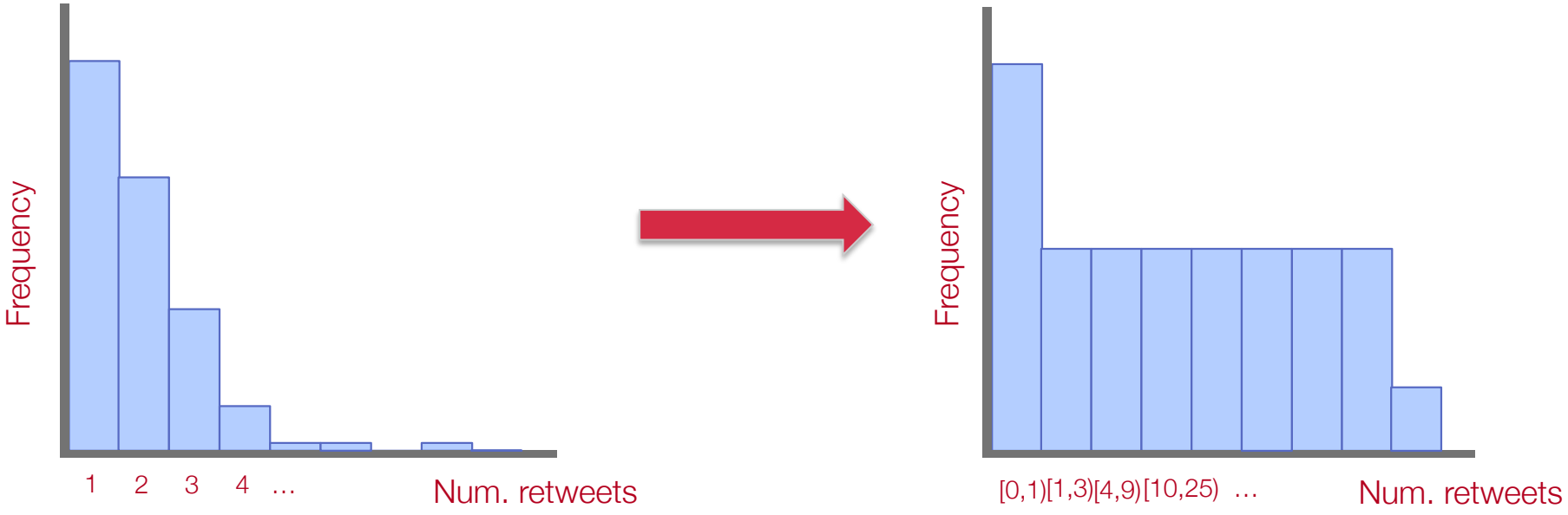

Training data
Tweet 1
f1, f2, f3, …
retweet count = 40
Tweet 2
f1, f2, f3, …
retweet count = 10
...
Unseen Tweet
f1, f2, f3, …
retweet count = ?
untrained
Bayesian Network
trained
Bayesian Network
Estimated
retweet count

Data


Main dataset
> 240,000 Tweets
90%
10%
Training dataset
Testing dataset
Model training
Feature extraction
Trained Bayesian Network
Test instances

‘Binning’ the retweet counts

To improve prediction performance



Interestingness 'scores'

Since predicting a number, can move from binary to continuous 'labels'

$$S_G(t) = \frac{\text{observed retweet count of } t}{\text{expected retweet count of } t\ *}$$

$$S_U(t) = \frac{\text{observed retweet count of } t}{\text{expected retweet count of } t\ **}$$

* Generated from the *global* model

** Generated from t's author's *user* model

Interestingness ‘scores’

$$S_G(t),\ S_U(t)$$

- > 1 t is ‘interesting’
- ≤ 1 t is ‘non-interesting’

Validating the *new* method

amazon mechanical turk
beta
Artificial Artificial Intelligence

Randomised controlled trial

750 Tweets tested.
Each tested…

… in 3 different ‘questions’

… by 3 different MTWs

Validating the *new* method

Select the Tweet(s) that you find the most interesting

VeryBritishProblems @SoVeryBritish
Loudly tapping your fingers at the cashpoint, to assure the queue that you've asked for money and the wait is out of your hands

VeryBritishProblems @SoVeryBritish
Running out of ways to say thanks when a succession of doors are held for you, having already deployed 'cheers', 'ta' and 'nice one'

VeryBritishProblems @SoVeryBritish
Looking into having your hands surgically removed after waving at someone who was waving at someone behind you

VeryBritishProblems @SoVeryBritish
Being unable to turn and walk in the opposite direction without first taking out your phone and frowning at it

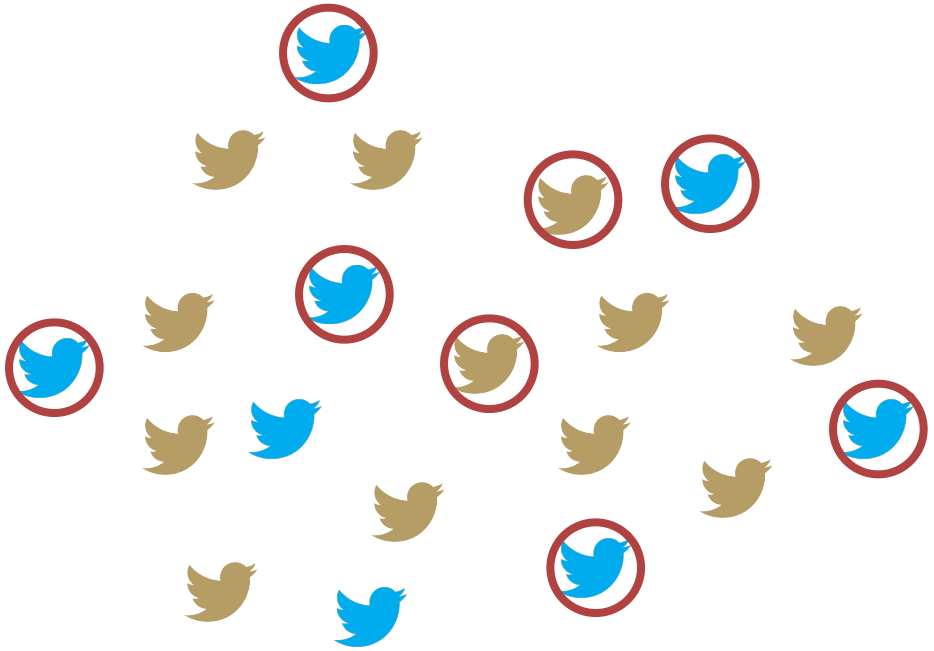
VeryBritishProblems @SoVeryBritish
Not quibbling with the unexpectedly high price, despite being certain your choices fully adhere to the rules of the Meal Deal

Validating the *new* method

325 / 450 answers with high confidence (> 66%)

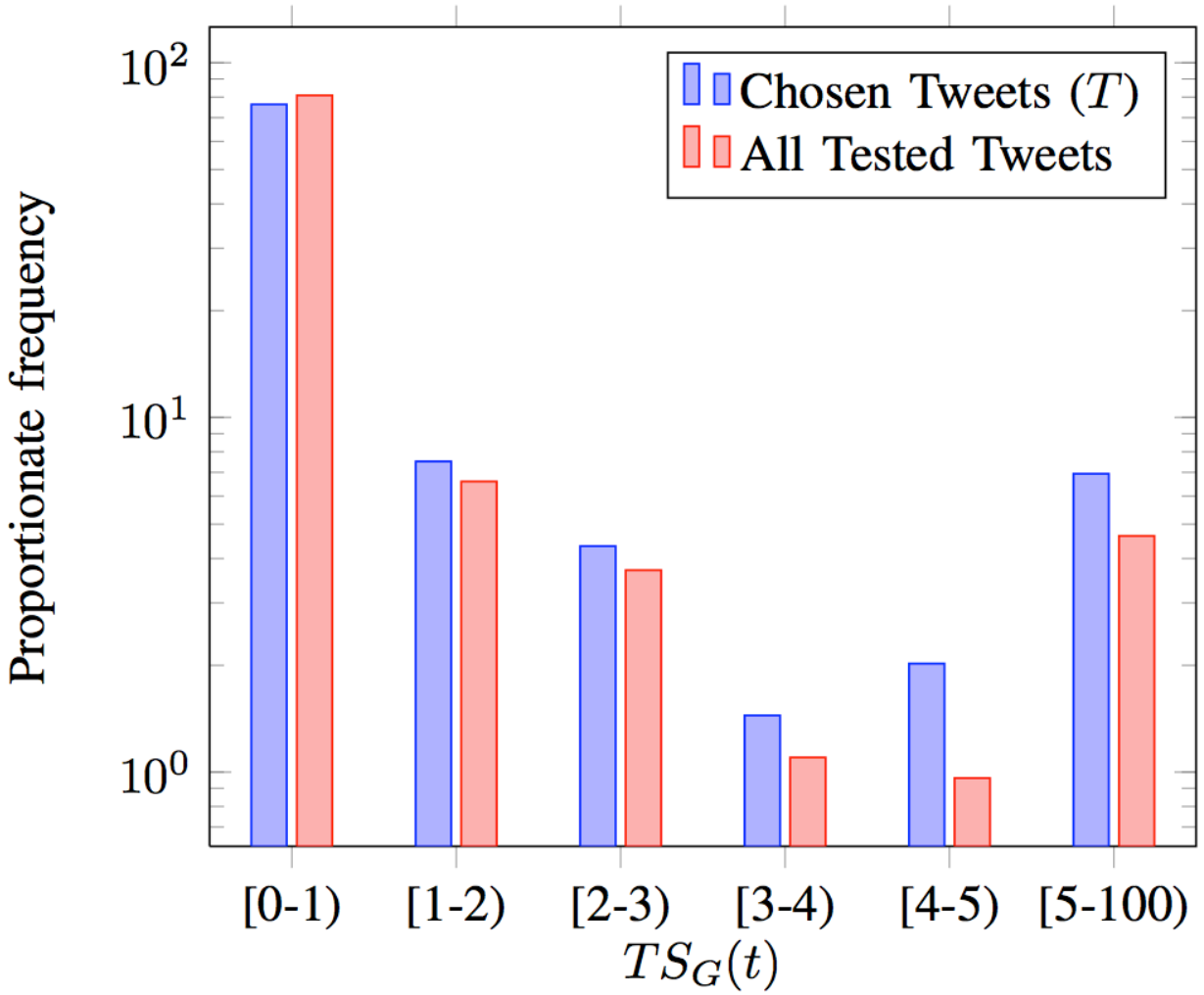
91 unique workers (diverse opinion)

65% agreement



Non-significant difference between scoring schemes

Validating the *new* method

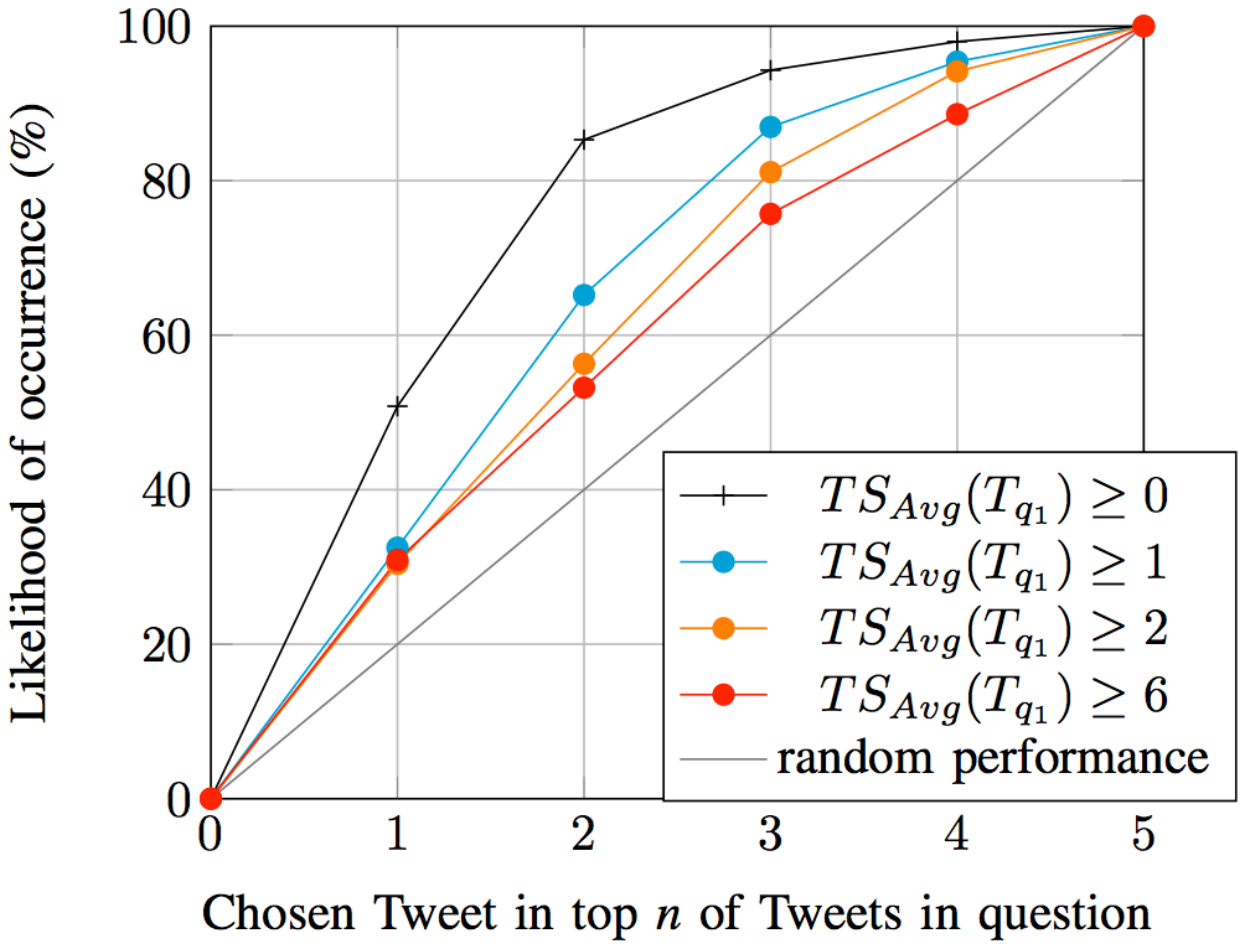


Chosen Tweets scores

>

Average scores

Validating the *new* method



70% chance of picking one of top 2 Tweets

What does this tell us?

Method is relatively good at detecting globally interesting information

What *doesn't* this tell us?

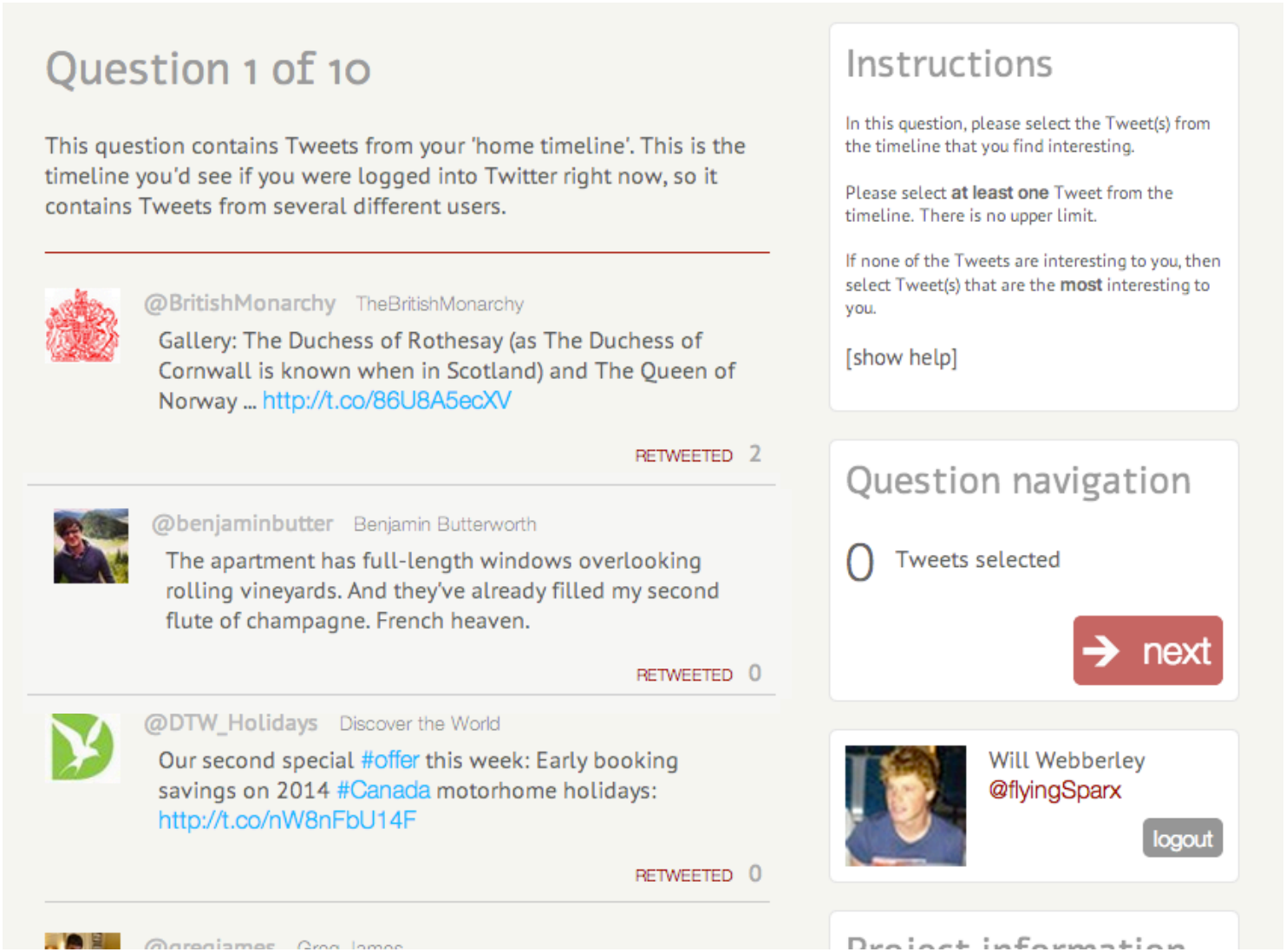
Is the method is able to pick out 'relevant' information for users?

Towards information ‘relevance’

Need users to assess Tweets within their own local network…

… from users they’ve already expressed an interest in through following

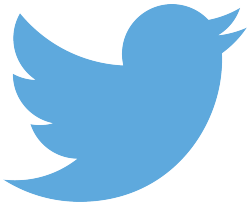
Towards information ‘relevance’



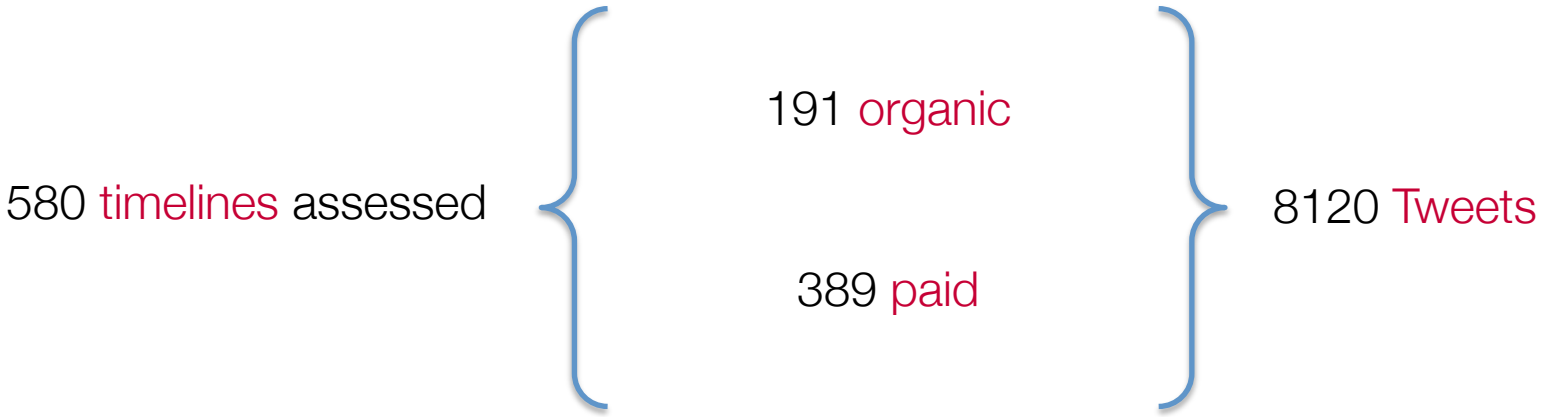
Towards information ‘relevance’

Paid participants

‘Organic’ participants

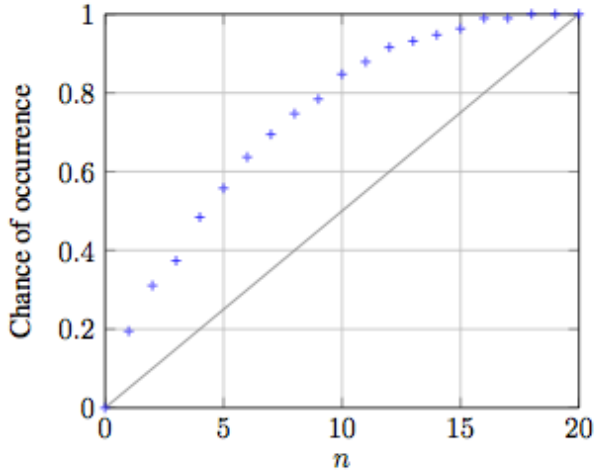


Towards information ‘relevance’

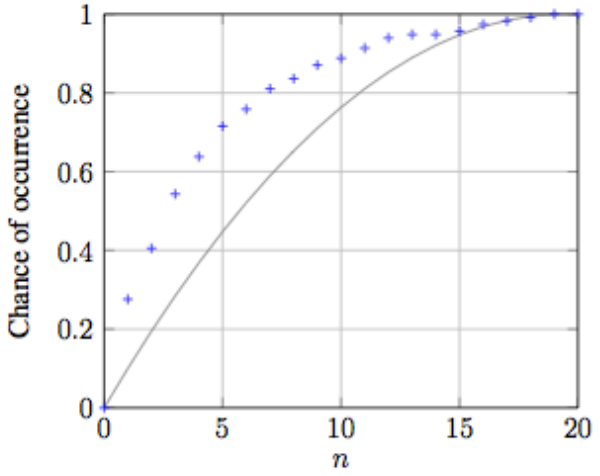


Non-significant difference between participant types

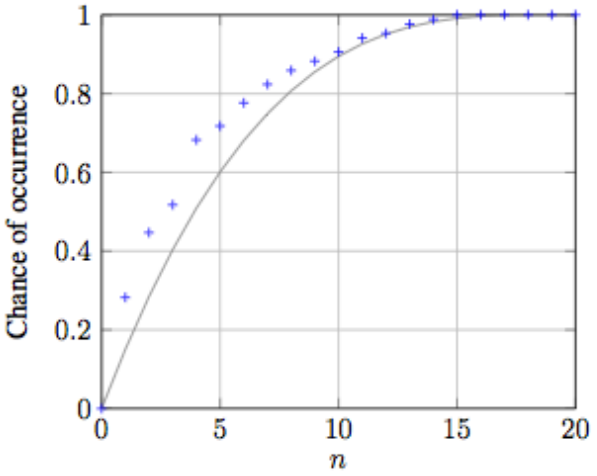
Towards information ‘relevance’



(a) In timelines where one Tweet was selected

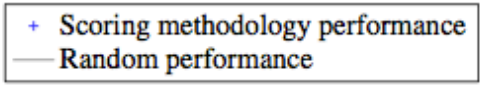


(b) In timelines where two Tweets were selected

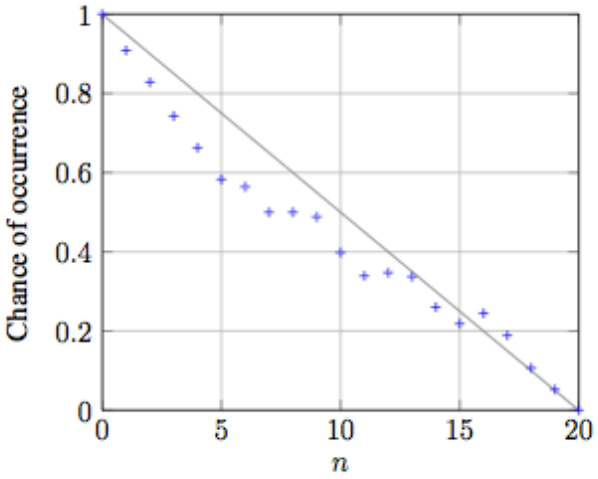


(c) In timelines where three Tweets were selected

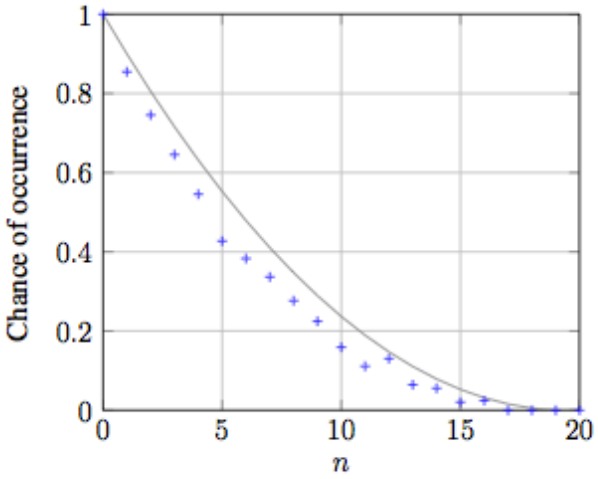
Selecting one of n *top* Tweets



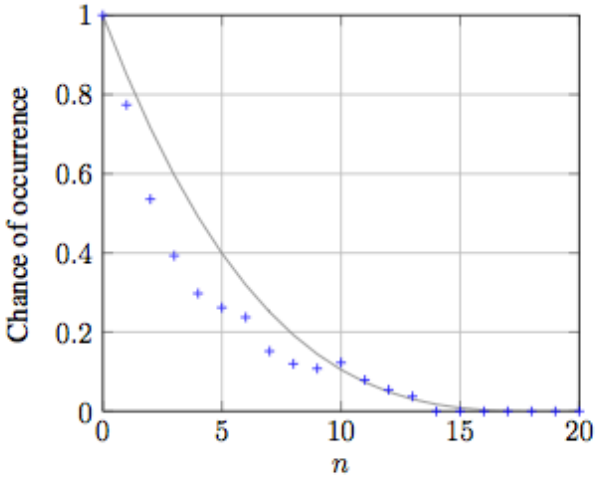
Towards information ‘relevance’



(a) In timelines where one Tweet was selected

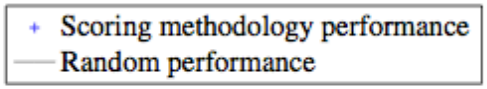


(b) In timelines where two Tweets were selected



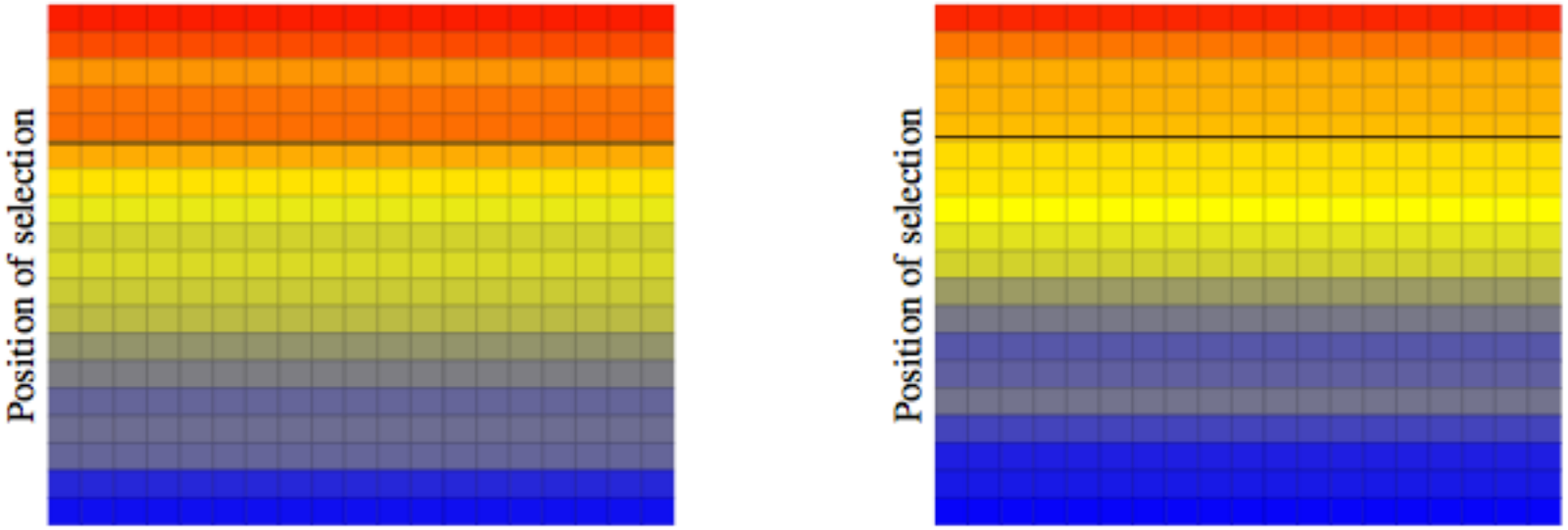
(c) In timelines where three Tweets were selected

Not selecting one of n ***bottom*** Tweets



Towards information ‘relevance’

Timeline position selections



(a) Selections made by organic participants

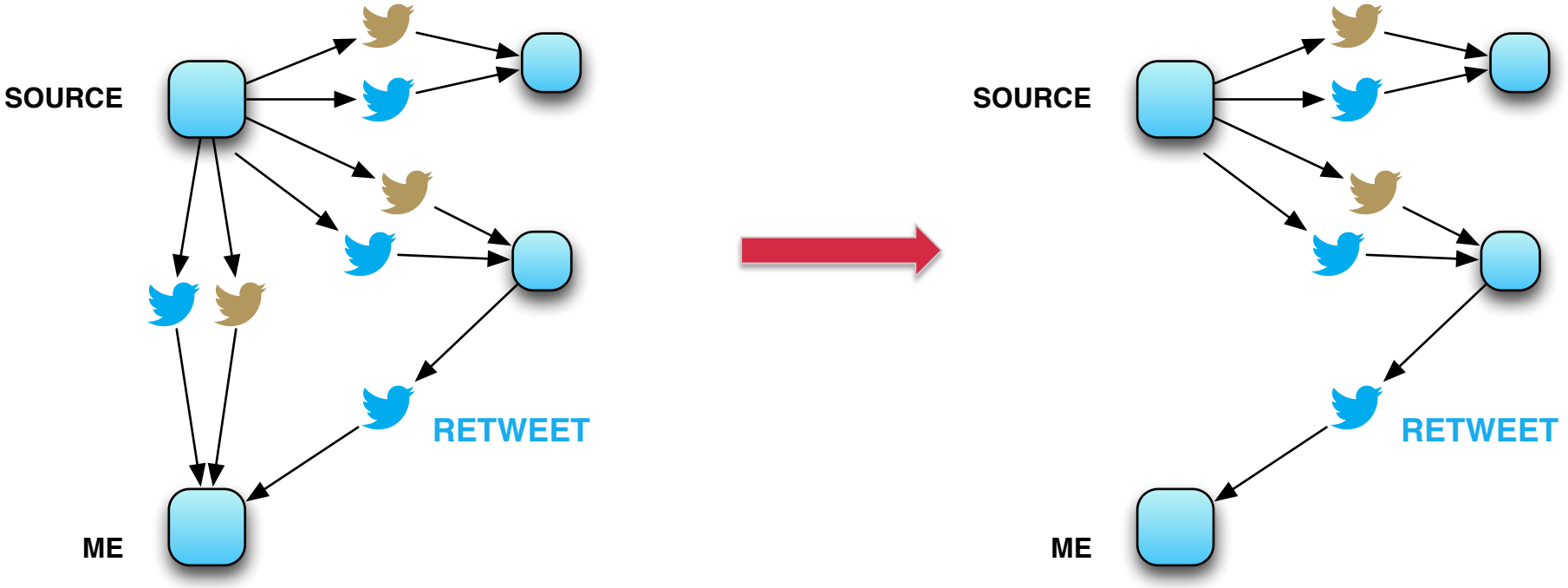
(b) Selections made by MTWs

Improvements

- Much more efficient
- More widely applicable
- More ‘on-demand’
- Supports Tweet ranking
- More accurate

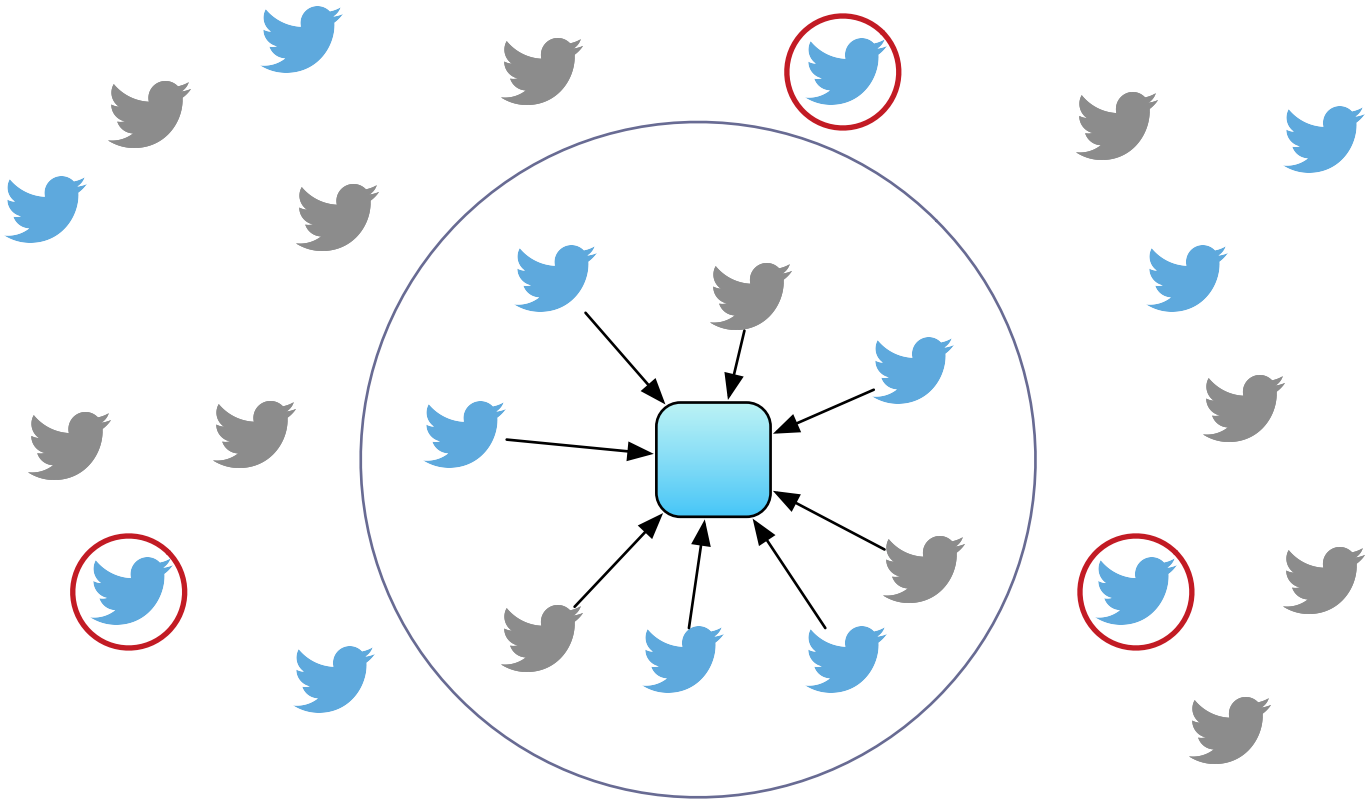
Taking this further…

Highlight superfluous edges in the social graph



Taking this further…

Building into concepts from others in the group



Thank you

Will M. Webberley

@flyingSparx

W.M.Webberley*

Stuart M. Allen

@monkeyfishgoat

Stuart.M.Allen*

Roger M. Whitaker

@profmobisoc

R.M.Whitaker*

* @cs.cardiff.ac.uk

School of Computer Science & Informatics

CARDIFF UNIVERSITY

PRIFYSGOL CAERDYDD