Investigating communities in the climate change debate on Twitter

James McMurray

This research investigates the presence and structure of communities of users who partake in the climate change debate on Twitter. Data was obtained via the Twitter API over 3 months on several hashtags, and strong community structure has been observed in some of these datasets. This demonstrates that there is an active debate on Twitter, and that like-minded users form communities. Possible future work and applications of machine learning will also be discussed. This research is primarily supervised by Dr. Hywel Williams of the Biosciences department of the University Of Exeter.

Supervisor: Dr. Hywel Williams
Biosciences
University Of Exeter

12/03/2013

UNIVERSITY OF EXETER

Outline of talk

- Introduction
  - Background
  - Aims
  - About Twitter
  - Setup
- Retweets
  - Retweets: Naive Retweet graphs
  - Retweets: Retweet chains
- Follower graphs
- Caveats
- Current work
  - Conversation graphs
  - Article/topic graphs
- Future work
  - Predictive modelling
  - Other sources
- Conclusion

Introduction: Background

- Preliminary research for broader work investigating behaviour-changing interactions with regard to climate change.
  - How do people change their opinions and behaviour, and how can this be utilised to change attitudes regarding climate change?
  - Is it possible to replicate success of anti-smoking, anti-drink driving campaigns?
  - What role can social media play in modern efforts to influence behaviour and opinions?
- Supervised by Dr. Hywel Williams in the Biosciences department, alongside Dr. Hugo Lambert of the Mathematics department and Dr. Timothy Kurz of the Psychology department (who is responsible for the funding).

Introduction: Aims

- What are the relative sizes of the sides of the climate change debate?
- What types of community structure exist amongst debaters?
  - Is it a polarised debate?
- Do people of differing views often interact, or do the users form "echo chambers"?
- Is the debate inclusive, or dominated by a few users?
- Who are the most important users?
- What role do media organisations have?
- Is it possible to observe evidence of "astroturfing"?

Possible community structure?

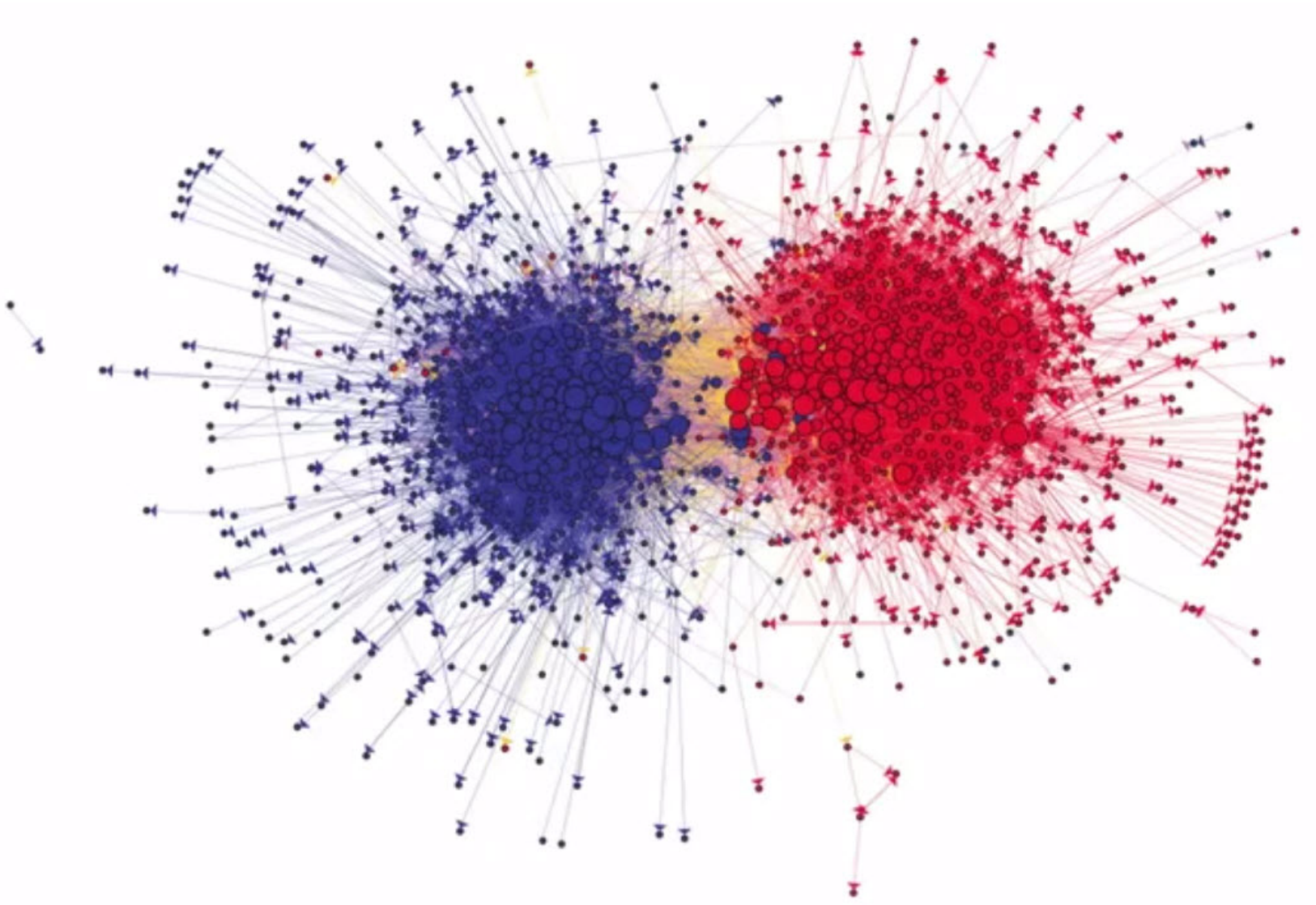


Figure : Example of community structure in American political blogs.
Source: Lara Adamic, Social Network Analysis course on Coursera

Introduction: About Twitter

- Users post tweets up to 140 characters in length.
- Users follow other users to receive their tweets on their homepage.
- Users may tag tweets with hashtags, which other users can search for - to follow tweets on a topic. E.g. discussing live TV shows, debates and so on.
- Users may retweet other user's tweets. This broadcasts that tweet to the user's followers, maintaining the original user as the source (note this is true even if one retweets a retweet, creating problems of lost information).
- Users may mention other users in their posts, to get their attention.
- Users can also post URLs which are automatically shortened (and also images and videos, for which hosting is provided).

Introduction: Setup

- Use the Twython module as an interface to the Twitter REST API.
- Use SQLite databases to hold the tweets and user information.
- Can obtain the following information from the API:
  - Tweet text
  - Timestamp, accurate to seconds
  - Username of posting user
  - Hashtags
  - Mentions
  - Username of first original source if it is a retweet
  - If the tweet is a reply to a user, that user's User ID and the ID of the tweet being replied to.
- Requests are rate limited to 180 requests per 15 minutes for tweet queries, but only 150 requests per 15 minutes for user queries (and for users with many followers, obtaining the full followers list can require multiple queries).

Introduction: Data

- We took data over 5 hashtags, and 2 quoted queries:
  - #climatechange 93,117 tweets
  - #globalwarming 37,816 tweets
  - #climate 114,909 tweets
  - #agw 7,137 tweets
  - #climaterealists 797 tweets
  - "Climate Change" 488,922 tweets
  - "Global Warming" 298,195 tweets
- Collecting data for approximately 3 months.
- For quoted queries, Twitter automatically groups related tweets.
- This appears to create bias: *Assessing the Bias in Communication Networks Sampled from Twitter*, González-Bailón *et al.* (2012), arXiv:1212.1684
- We have focussed on hashtag data: #climatechange, #globalwarming and #agw specifically.

Outline

- Introduction
  - Background
  - Aims
  - About Twitter
  - Setup
- Retweets
  - Retweets: Naive Retweet graphs
  - Retweets: Retweet chains
- Follower graphs
- Caveats
- Current work
  - Conversation graphs
  - Article/topic graphs
- Future work
  - Predictive modelling
  - Other sources
- Conclusion

Retweets

- Started the investigation focussing on retweets.
- Because these prove that the original tweet was read, and that the reader felt strongly enough to re-broadcast the tweet.
- However, while we are provided the original source of the retweet, the intermediate sources are lost.
- But can plot a weighted, directed, "naive" retweet graph, connecting retweeters to the original source easily.

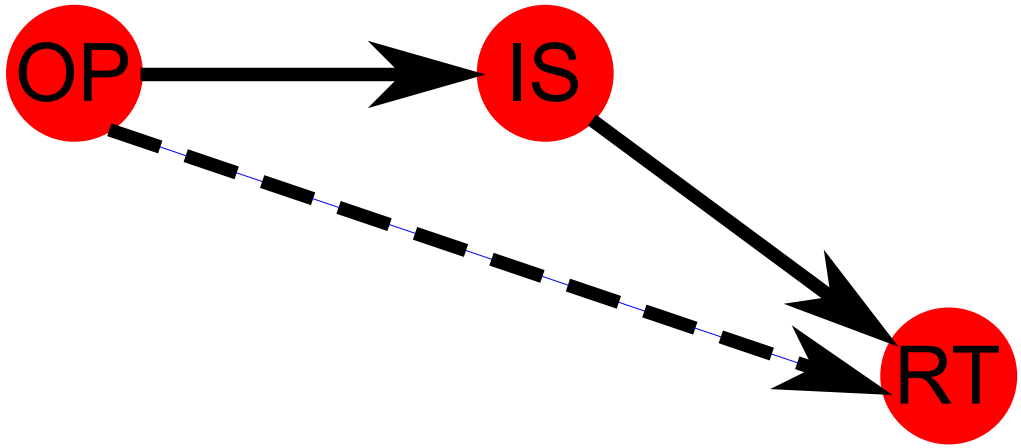
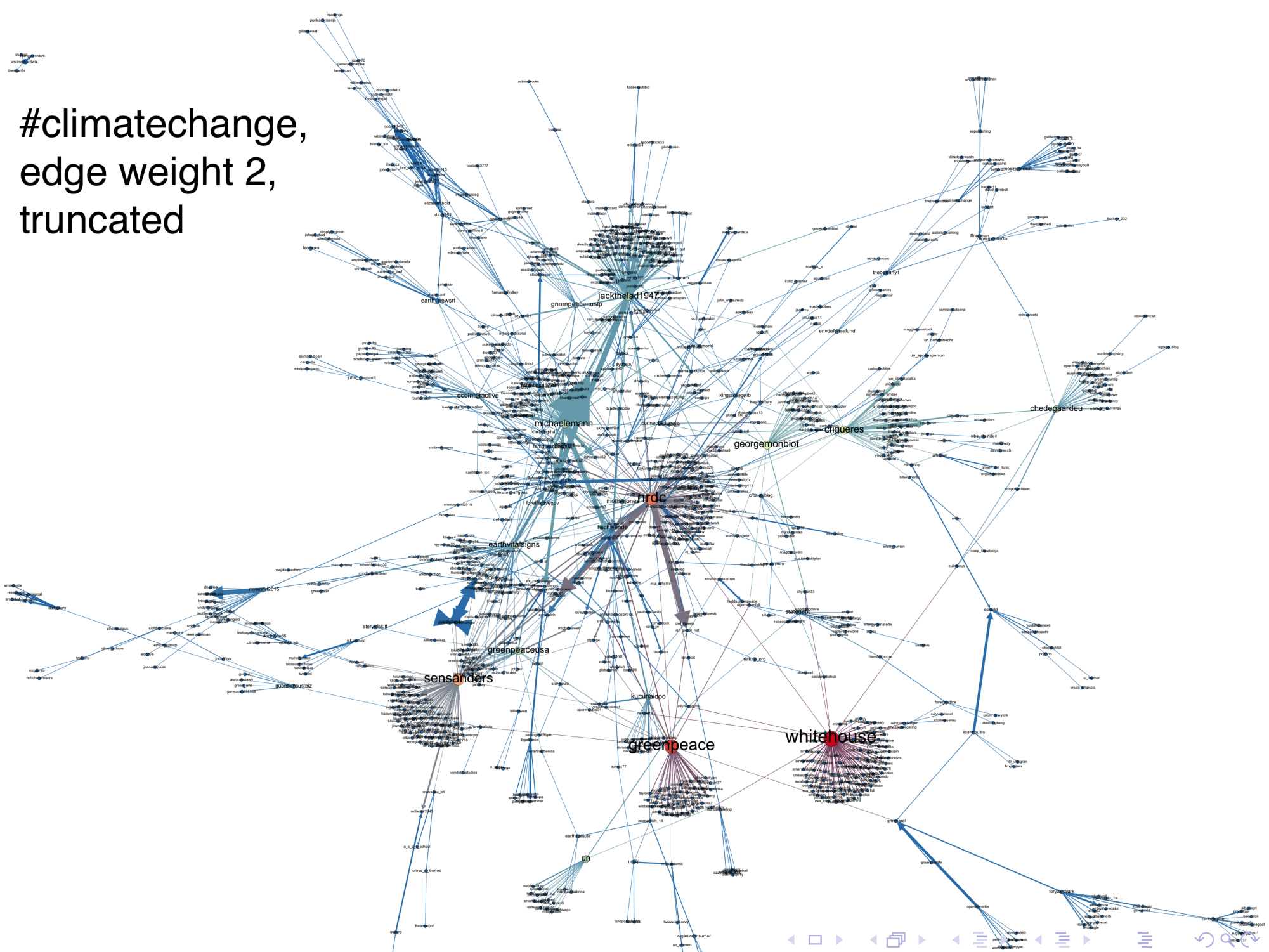
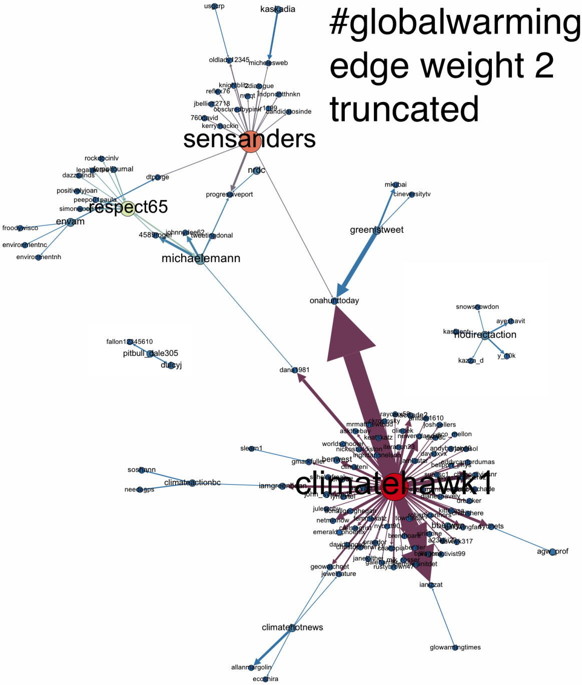


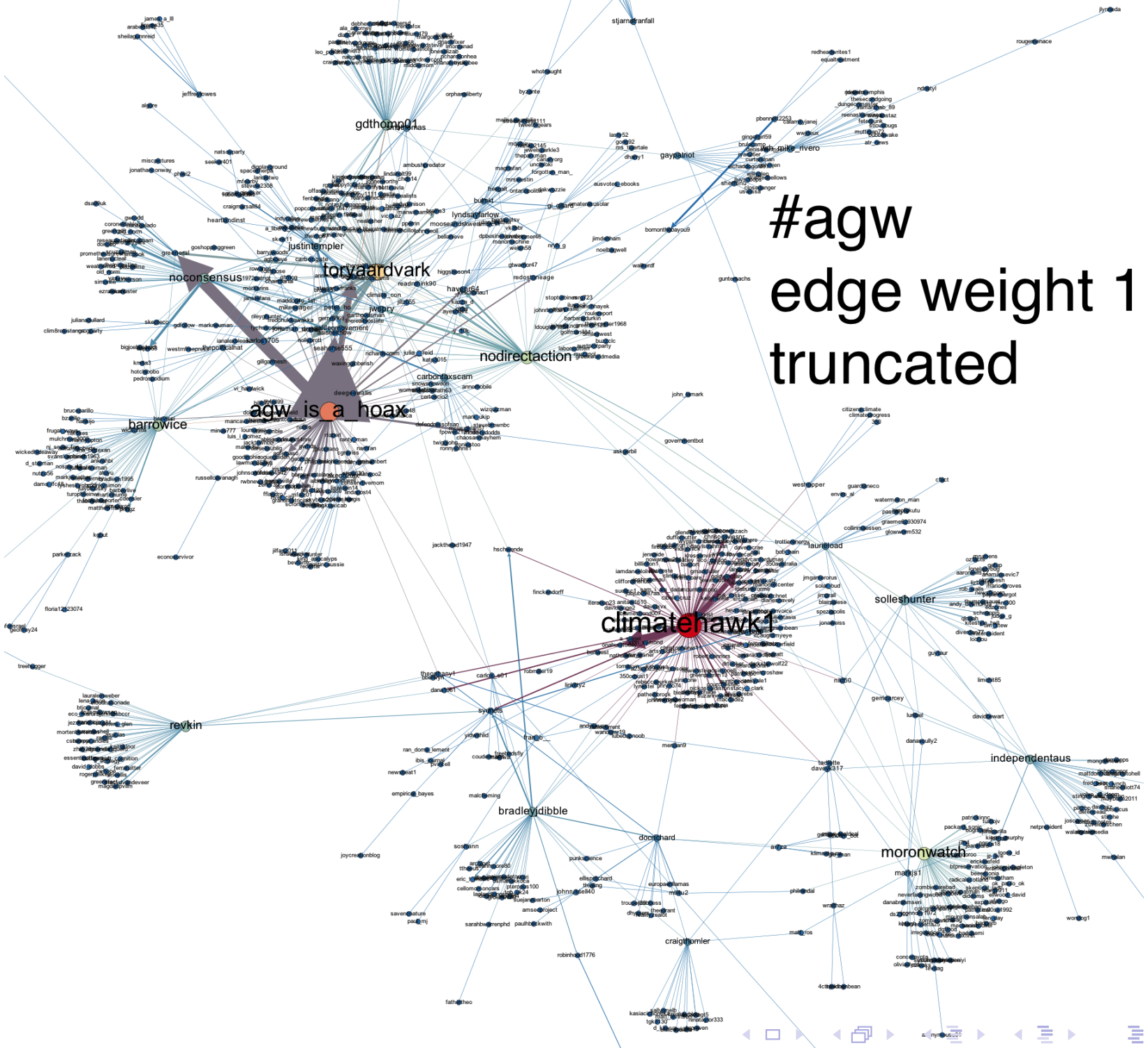
Figure : Diagram demonstrating loss of information. Retweet may have occurred intermediates source (full lines), but we can only observe dashed line to original source.

#climatechange, edge weight 2, truncated



#globalwarming edge weight 2 truncated



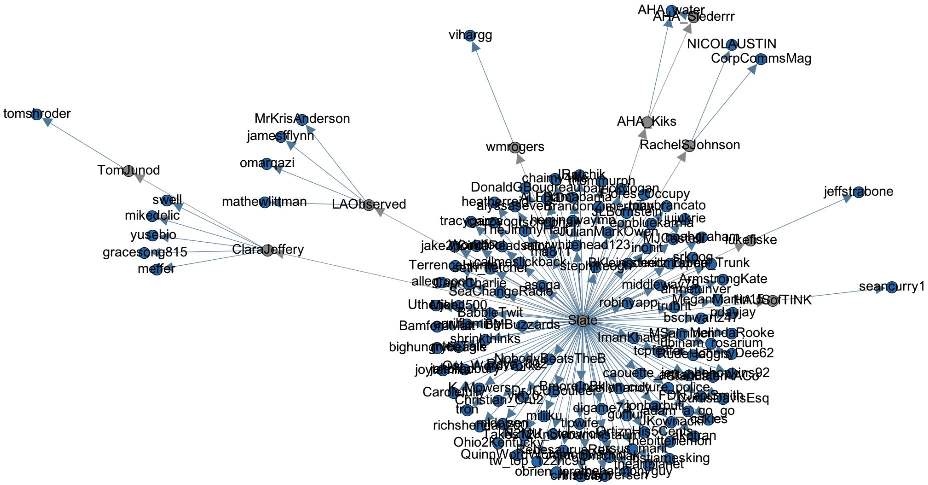

#agw
edge weight 1
truncated
agw_is_a_hoax
climatehawk1
nodirectaction
torvaardvark
noconsensus
barrowice
revkin
bradleyjdibble
moronwatch
laurelroad
solleshunter
independentaus
gdthomm01
justintempler
lyndsaydarlow
mike_rivero
gaypalriot
havenot64
jwspry
carbontaxscam
craigthomler
docithard
markjs1

Retweets: What do we see on Naive RT graphs?

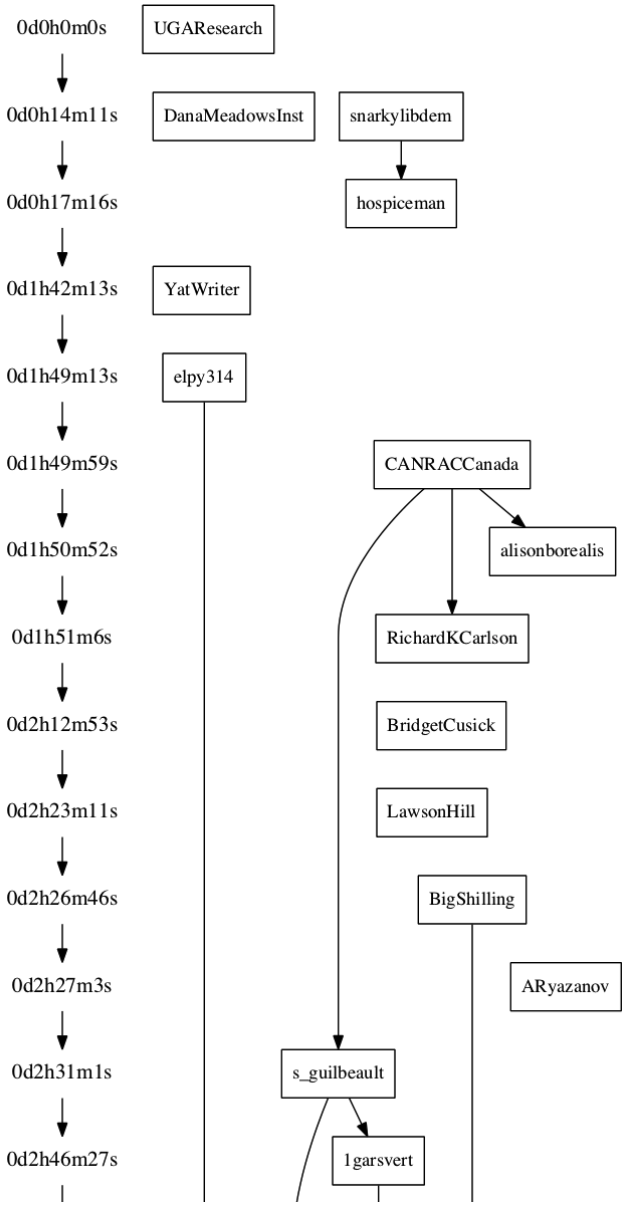
- No apparent community structure in #climatechange data. Dominated by news reports and some activists.
- Less traffic on #globalwarming, although small disconnected clusters of skeptics and music fans are visible alongside the activists.
- Community structure visible on #agw - it appears this is where the majority of the debate is.
- Hub structure - but this is implicit due to the collapsing of the retweet chains.
- Is this method appropriate?

Rebuilding the retweet chains

- We can attempt to rebuild the retweet chains if we use a few assumptions:
  - If the user follows the original source, then assume they retweeted directly from the original source.
  - If the user does not follow the original source, check which users they are following who retweeted the same tweet (from the same original source) before them. Create edges to these users to try to reconstruct path backwards.
  - If we cannot find a viable path, then just add an edge directly to the original source (perhaps they found the tweet by other means, such as direct search).
- This assumes users have not used the search to find these tweets, and are instead following the user.
- This has had a good non-failure rate with our datasets, but would be inappropriate in a case of a lot of search on hashtags, such as in a trending hashtag or discussing major news/television broadcast.


tomshroder
TomJunod
MrKrisAnderson
jamesfflynn
omarqazi
mathewlittman
LAObserved
swell
mikedelic
yusebio
gracesong815
meffer
ClaraJeffery
viharggg
wmrogers
AHA_water
AHA_Slederrr
NICOLAUSTIN
CorpCommsMag
AHA_Kiks
RachelSJohnson
jeffstrabone
lukefiske
seancurry1
Slate
SeaChangeRadio
BabbleTwit
shrinkthinks
NobodyBeatsTheB
ArmstrongKate
robinyapp
bschwartz47
MelindaRooke
rosarium
Dee62
caouette
police
CurtisDavisEsq
a_go_go
JKowmacki
thebitterlemon
QuinnWord
Ohio2Kentucky
richshenton
tron
Christian_Cruz2
allegroach
Terrence
tracycarroll
heatherreid
DonaldGBoudreau
chaim
inonit
srkoog
MJCastro
tcpfeifer
ImanKhald
Trunk

Retweets: Plot of retweets against time



- From discussion of an article about Climate Change published in Slate magazine
- Provides timeline of all tweets of article
- Full graph is much larger
- What do we get on average?

UNIVERSITY OF EXETER

Retweets: Calculations over retweet chains I

- Ran calculations to find the distribution of chain lengths in the timeline graphs, over all the unique retweets.
- Note this takes a long time however, due to the need to collect almost all of the user data, which is very rate limited.
- Unfortunately Twitter drops URL information if retweets are truncated, so it was necessary to just test for existence of the middle 80 characters to check if tweets were of the same topic/article.
- For 14,062 unique "topics" the distribution of chain length was:
  0: 232, 1: 15027, 2: 282, 3: 47, 4: 49, 5: 31, 6: 17, 7: 4

Retweets: Calculations over retweet chains II

- For 14,062 unique "topics" the distribution of chain length was:
  0: 232, **1: 15027**, 2: 282, 3: 47, 4: 49, 5: 31, 6: 17, 7: 4
- Strongly peaked at 1, so naive retweet graphs seem a reasonable approximation to use (to avoid polling user data).
- This is expected, since it doesn't cost anything to follow a user, and if you retweet a user you are likely to follow them (this could also be tested).
- The higher chain lengths were mostly retweets by media companies through their subsidiaries. A chain length of zero means the tweet was not retweeted (although we have biased against this by only considering topics (i.e. tweets containing the substring) which were retweeted at least once, while most tweets are never retweeted.)

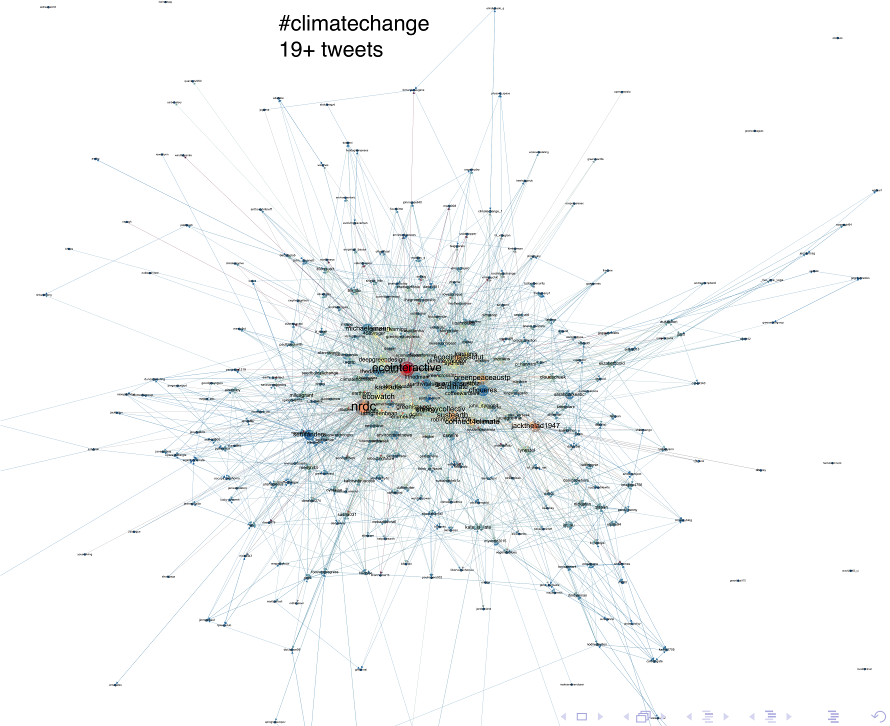
Outline

- Introduction
  - Background
  - Aims
  - About Twitter
  - Setup
- Retweets
  - Retweets: Naive Retweet graphs
  - Retweets: Retweet chains
- Follower graphs
- Caveats
- Current work
  - Conversation graphs
  - Article/topic graphs
- Future work
  - Predictive modelling
  - Other sources
- Conclusion

Follower graphs: Who follows who?

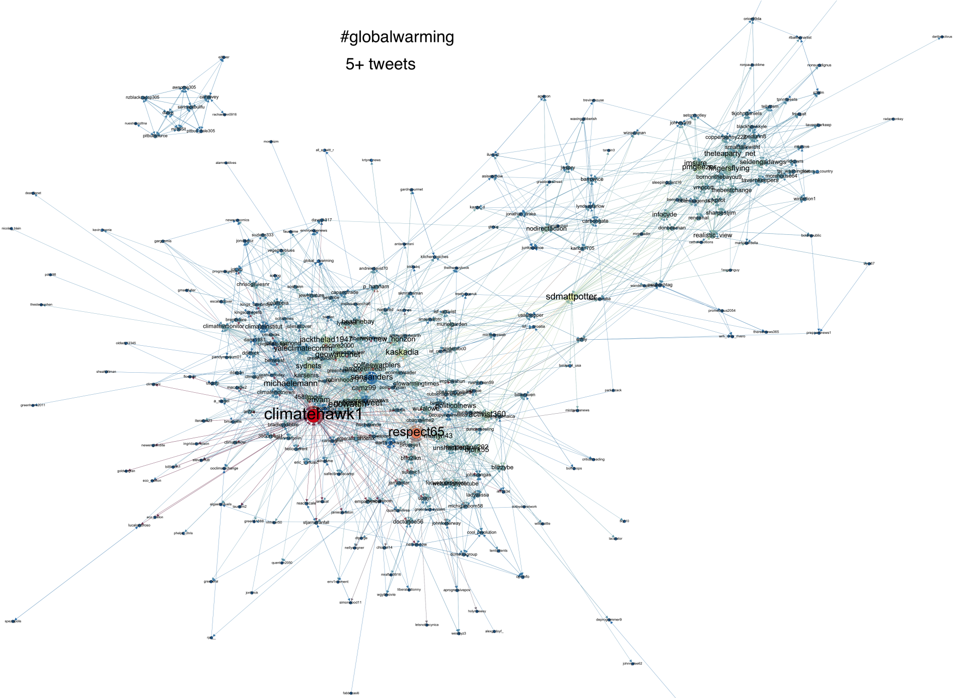
- From the user data we can also plot follower graphs for the most prolific users in each dataset (chose approximately top 400 users).
- We observe much stronger community structure in these graphs.
- Like minded users tend to follow each other.
- Abundance of edges makes clustering more effective compared to naiveRT plot.
- Clustering achieved via the Proportional Yifan-Hu algorithm in Gephi which is fundamentally a proportional force-directed algorithm.
- Paper available at: `http://www2.research.att.com/~yifanhu/PUB/graph_draw_small.pdf`

#climatechange
19+ tweets



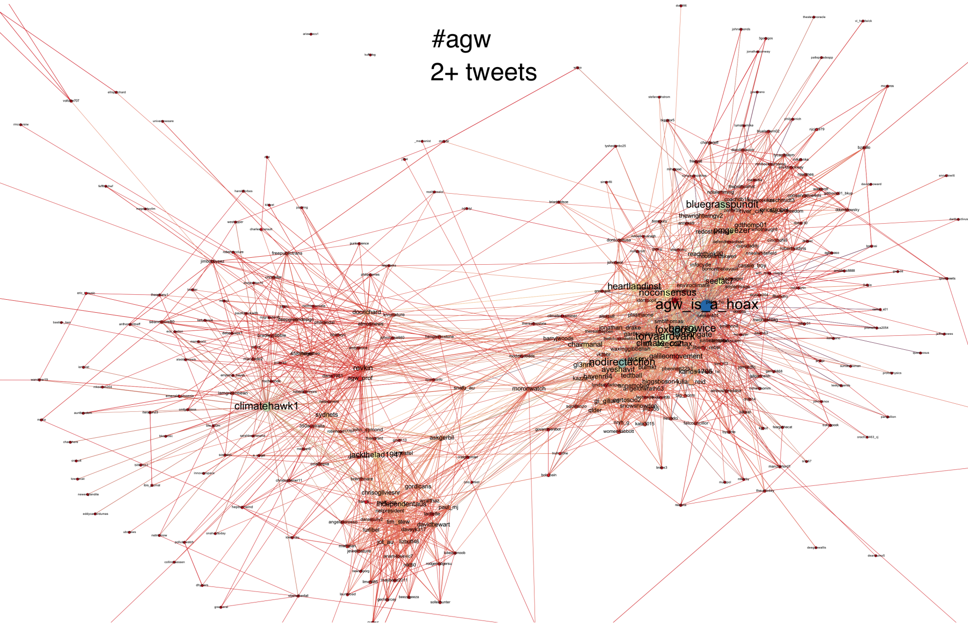
#globalwarming

5+ tweets



#agw

2+ tweets



agw_is_a_hoax
bluegrasspundit
heartlandinst
noconsensus
toryaardvark
nodirectaction
climatehawk1
revkin
agw_prof
jackthelad1947
independentaus
chairmanal
moronwatch
askgerbil
gordicans
sydnets
docrichard
atmosnews
ayeshavit
galileomovement
thewrightwingv2

Follower graphs: What do we see

- Strong community structure in #agw and #globalwarming, less in #climatechange.
- Perhaps there is a difference between hashtags primarily used for news dissemination and those used for debate.
- Work is undergoing to manually judge users as skeptics, activists or other.
- Need to verify that clusters are ordered on lines of the debate and not other causes (i.e. one group discussing wind power and another discussing nuclear power)
- Though it appears that the clusters are split on their support/opposition to the theory of anthropogenic global warming.

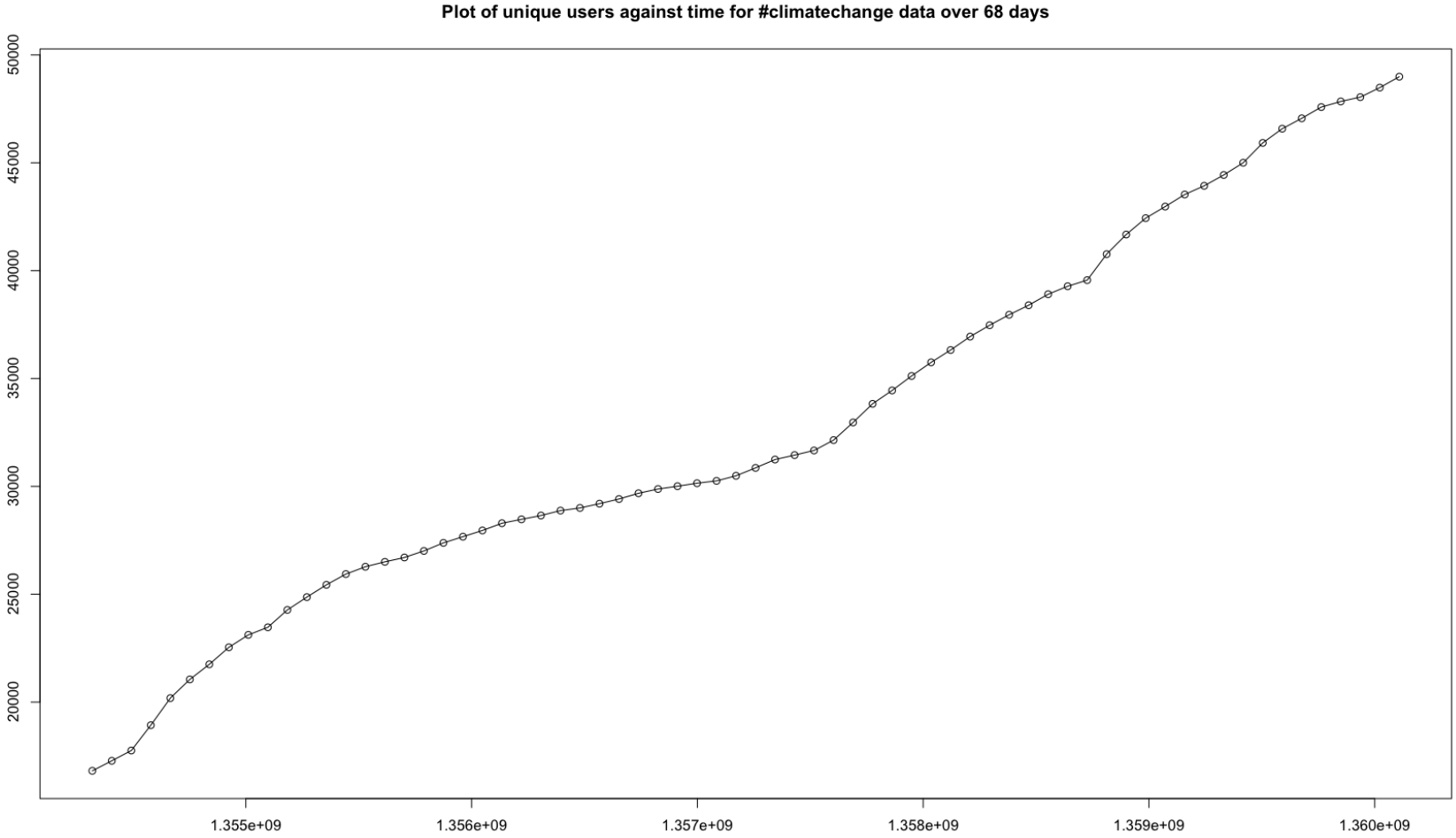
Outline

- Introduction
  - Background
  - Aims
  - About Twitter
  - Setup
- Retweets
  - Retweets: Naive Retweet graphs
  - Retweets: Retweet chains
- Follower graphs
- Caveats
- Current work
  - Conversation graphs
  - Article/topic graphs
- Future work
  - Predictive modelling
  - Other sources
- Conclusion

Caveats: The need for assumptions

- As demonstrated in rebuilding the retweet chains, sometimes extracting useful information requires making assumptions.
- We look only at hashtags - many users rarely or never use hashtags.
- The network varies over time, for example we cannot access user information for banned users, so we have removed these from the dataset.
- We only see follower information at the present time, not a snapshot of it at the time of the tweets.
- This could profoundly affect the retweet chains since a user is likely to start following someone who they have retweeted.
- Likely has an effect but is hard to directly measure.
- How do we know when we have captured a representative image of the debate?

Caveats: Plot of unique users against time - #climatechange



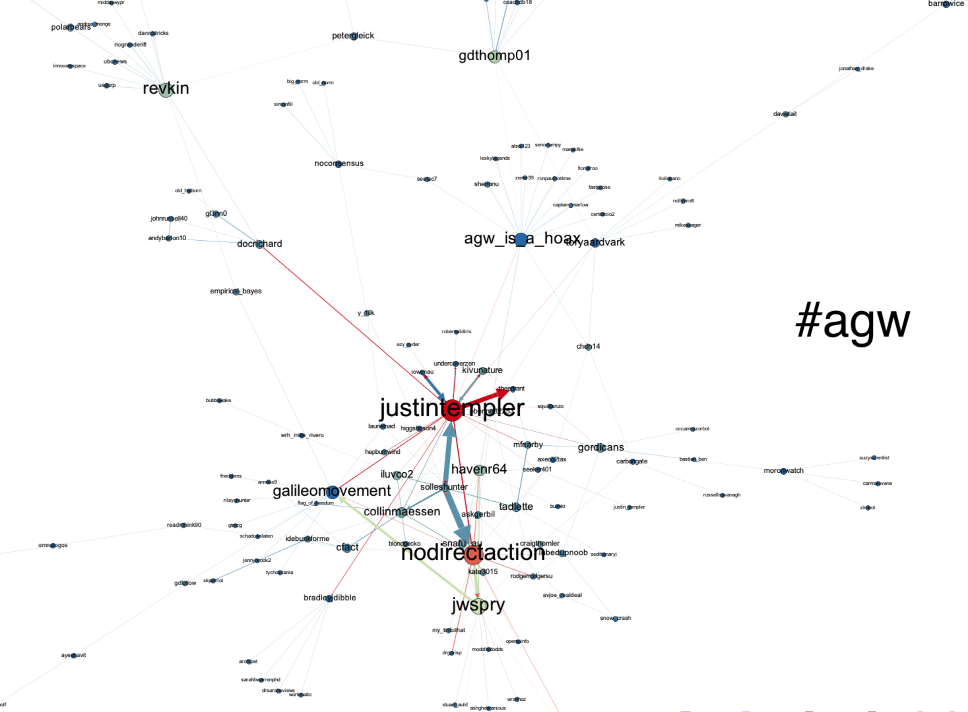
Outline

- Introduction
  - Background
  - Aims
  - About Twitter
  - Setup
- Retweets
  - Retweets: Naive Retweet graphs
  - Retweets: Retweet chains
- Follower graphs
- Caveats
- Current work
  - Conversation graphs
  - Article/topic graphs
- Future work
  - Predictive modelling
  - Other sources
- Conclusion

Current work: Conversation graphs

- We can also produce directed, weighted graphs of mentions between users.
- Alongside the information from the API about replies, this provides a comprehensive image of the direct interactions between users.
- From this we could look at how direct debates are started and propagated.
- Which users are more likely to start a debate?
- Have produced first graph, but need more data.

#agw



revkin
gdthomp01
petergleick
polarbears
noconsensus
docrichard
agw_is_a_hoax
toryaardvark
justintempler
nodirectaction
jwspry
galileomovement
havenr64
collinmaessen
iluvco2
solleshunter
tadlette
askgerbil
cfact
gordicans
mfearby
kivunature
bradleyjdibble
ideburnforme
empirical_bayes
chdn14
davetait
moronwatch
carbongate
kate3015

Current work: Sentiment analysis

- Can we use some form of sentiment analysis to judge whether users are skeptics or activists?
- Initially looked at implementing Naive Bayes classifier, but this is not a simple case:
  - No clear identifying words.
  - Lots of sarcasm.
- Perhaps more sophisticated Natural Language Processing methods could overcome these problems.
- Is it possible to judge users as skeptics or activists based solely on their tweets?
- Other research has focussed on sentiment analysis for Twitter, such as *Credibility in Context: An Analysis of Feature Distributions in Twitter* by O'Donovan *et al.*, PASSAT (2012)

Current work: Article/topic graphs

- We can also produce a weighted, undirected graph, with edges between users who have posted the same article/tweet - regardless of whether they follow each other.
- Abundance of edges should make community structure more clear.
- Will communities here line up with communities in the follower graphs?
- Work still underway.
- Could also look at this from a Machine Learning perspective - unsupervised learning based on articles posted for each user.
- Would be interesting to compare approaches.

Outline

- Introduction
  - Background
  - Aims
  - About Twitter
  - Setup
- Retweets
  - Retweets: Naive Retweet graphs
  - Retweets: Retweet chains
- Follower graphs
- Caveats
- Current work
  - Conversation graphs
  - Article/topic graphs
- Future work
  - Predictive modelling
  - Other sources
- Conclusion

Future work: Predictive modelling

- Can we use the parameters of the network to predict its behaviour in the future?
- Model transmission of information across the network
  - How does the polarisation effect this?
  - Are they true echo chambers?
- Can we discover necessary conditions for popular articles?
- Can we predict articles that will become popular?
  - Can use Streaming API for real time access but "random" sample is limited.
- Other research has shown success. Master's thesis: *Trend or No Trend: A Novel Nonparametric Method for Classifying Time Series* by S. Nikolov at MIT - 95% true positive rate and a 4% false positive rate using supervised learning with a latent source model.

Future work: Other sources

- Could also look at popular blogging platforms: Blogger, Wordpress, Tumblr.
  - Perhaps match users with Twitter users in some cases.
  - May be hard to extract nature of interaction, as much is in text.
- Apply to news aggregation websites: Reddit, Slashdot, Digg
  - May have opportunity to see exactly what stories users have up-voted and down-voted.
  - Can then look for concerted actions and communities.
  - Is there evidence of "astro-turfing"?

Outline

- Introduction
  - Background
  - Aims
  - About Twitter
  - Setup
- Retweets
  - Retweets: Naive Retweet graphs
  - Retweets: Retweet chains
- Follower graphs
- Caveats
- Current work
  - Conversation graphs
  - Article/topic graphs
- Future work
  - Predictive modelling
  - Other sources
- Conclusion

Conclusion

- Work is still underway, but there appears to be strong communities in the climate change debate on Twitter.
- We are still to investigate the results of this on information transmission in the network.
- Popularity of social media websites provides a goldmine of public data.
- Allows new efforts in quantitative sociology, applications in marketing, politics, etc. - any information transmission.
- Machine learning has a very important role in analysing the data and producing predictive models.

Thanks for your time.

Questions?