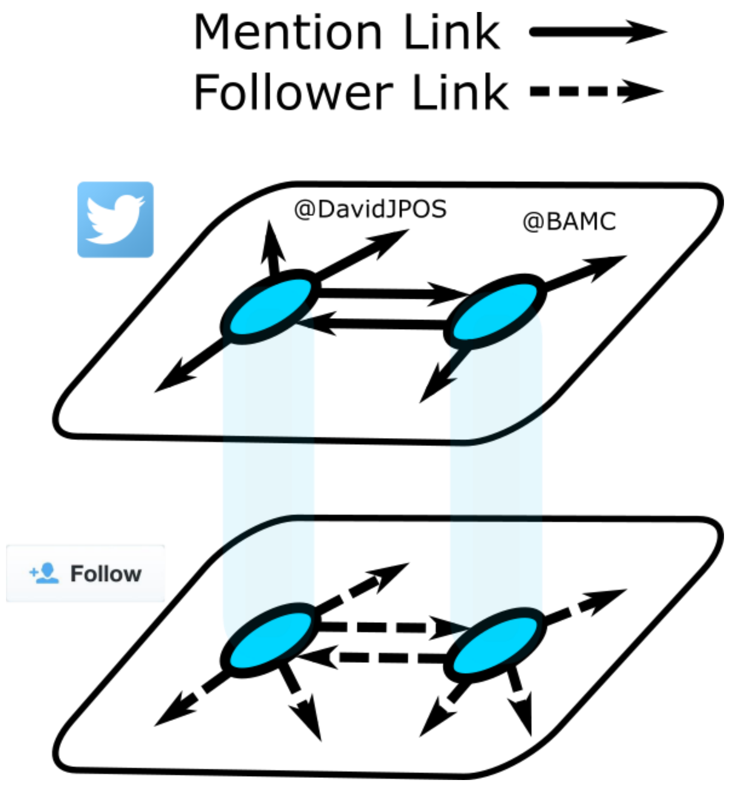
CREATING NETWORK FROM DATA

Networks from data

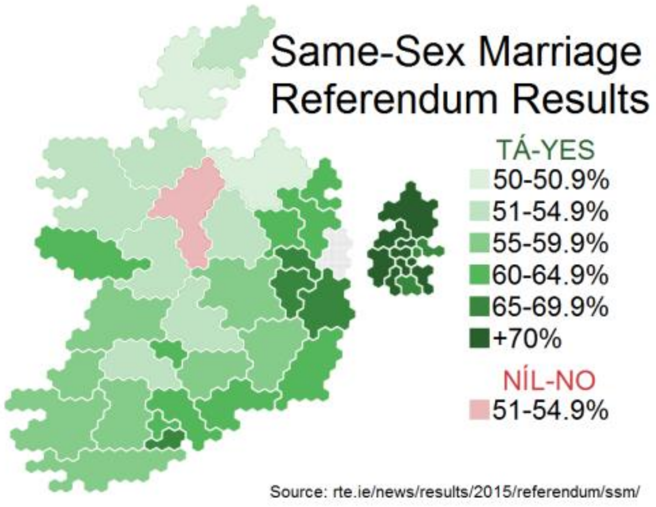
- Using similar data recreate similar analysis as a paper
  - Create a network
  - Calculate the properties of the network (we already have a little experience in this!)
  - Discuss properties like homophily and test for them
  - How generate random networks

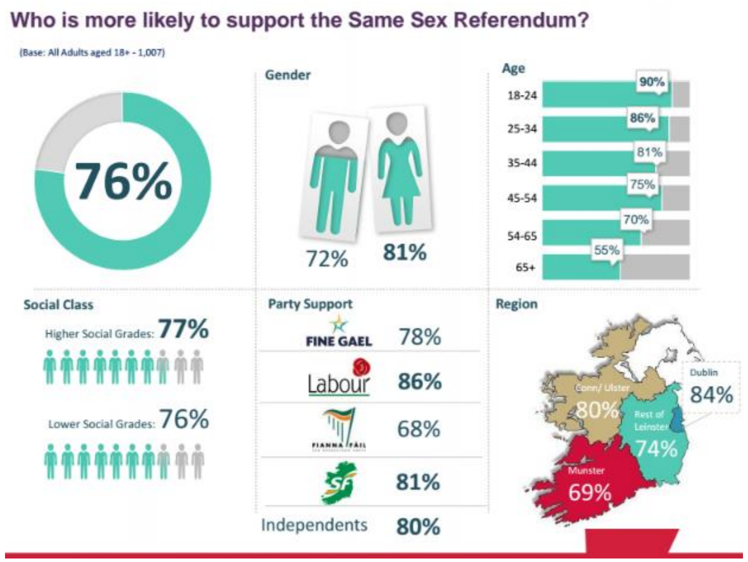


Irish Marriage Referendum Data

- Irish Marriage referendum
  - 22nd of May 2015
  - Passed by a 62% majority
  - High voter turn out 60%

- Collected an extensive dataset
  - “#marriageref” & “#marref”
  - 7th and the 23rd of May
  - 144,077 users & 499,642 tweets

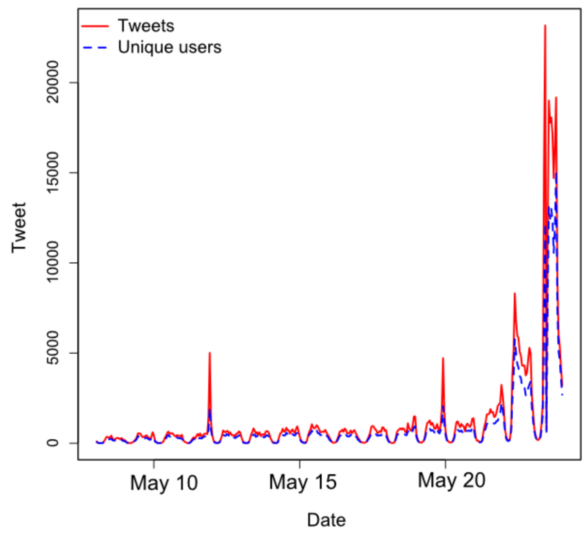




Irish Marriage Referendum Data

- Each tweet contained
  - 20 variables
    - Screen name
    - Time stamp
    - Text
    - Geolocation, etc
- For each user
  - 22 variables
    - Screen name
    - Description, etc
- Friends list
  - 177,669,550 directed links between users

- First step?



Irish Marriage Referendum Data

- Just a work of warning
  - A lot of people have tried to predict stuff with twitter data…

> **"I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper" -- A Balanced Survey on Election Prediction using Twitter Data**
>
> Daniel Gayo-Avello
>
> *(Submitted on 28 Apr 2012)*

- Why?

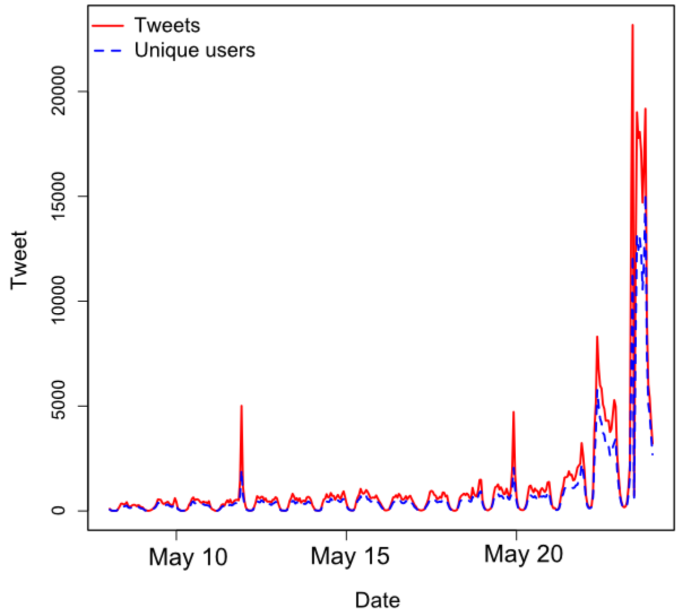
- Twitters API's give a sample
  - And worse again probably not random sample
  - Demographically not representative
  - Geographically not representative
  - Activity rates between users differ... A lot!

Metric?

- Volume of tweets?
- Users accounts?
- What about bots?
- Etc etc etc.

Irish Marriage Referendum Data

- But with those in mind…
- Interested in sentiment on networks
  - How positive or negative content/users are
- Does sentiment matter?
  - Sentiment sent and received?
  - Does sentiment cluster between users
    - Proxy for homophily?
  - Is it useful for classification of voters
    - Can we find yes and no voters?



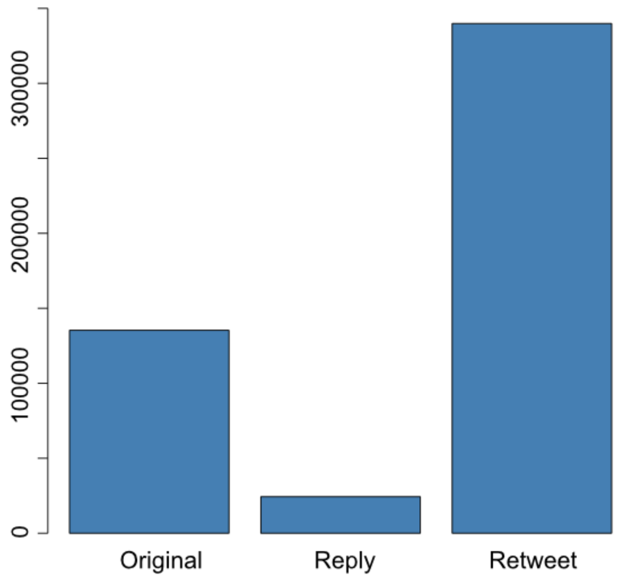
Data collection

- #marref – very popular hashtag
- #marriageref
  - not very popular (397)

- Types of tweets
  - Of the 499,642
    - Original 135,370 (27%)
    - Reply 24,397 (5%)
    - Retweet 339,875 (86%)

- Next step, homophily (sentiment)

	Frequency
#marref	499635
#voteyes	56299
#yesequality	44795
#hometovote	18761
#ireland	13661
#yes	13242
#voteno	11773

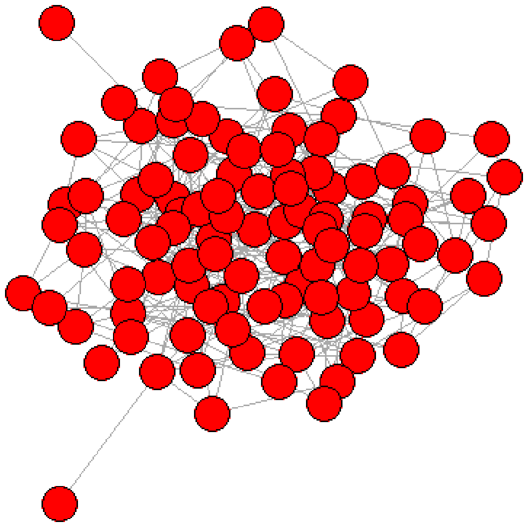


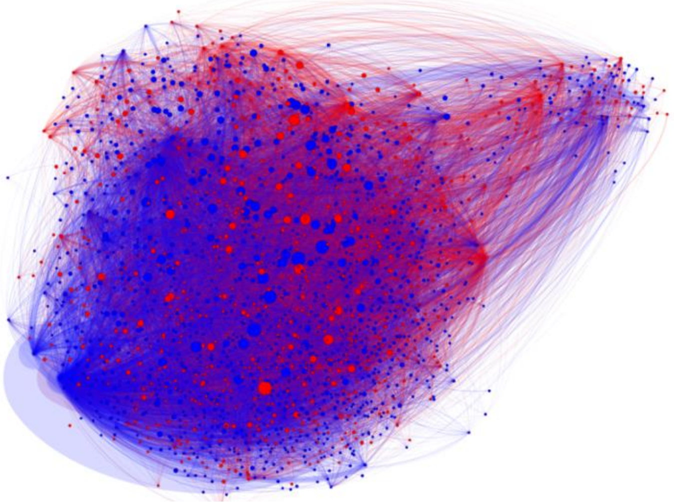
HOMOPHILY

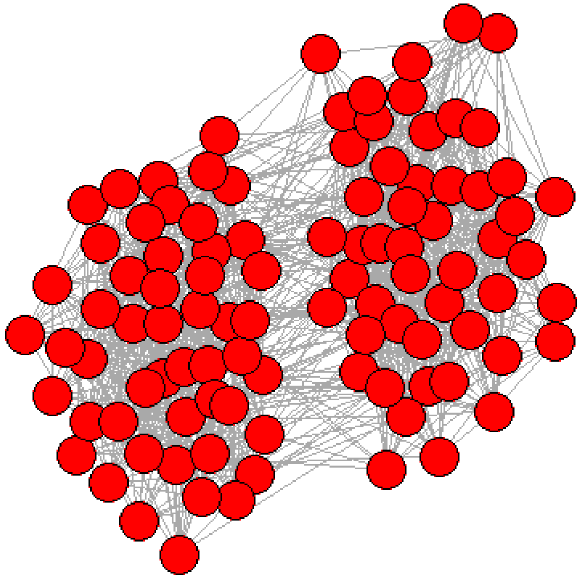
Homophily, what is it?

- ‘Birds of a feather flock together’
  - Do you see any homophily?

- What drives homophily
  - For music preferences?
  - For sport team?

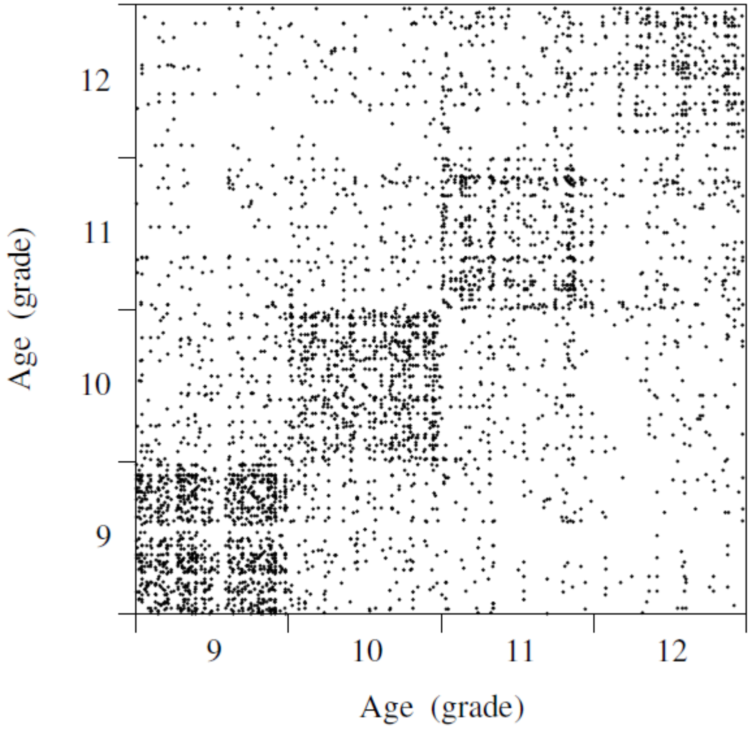






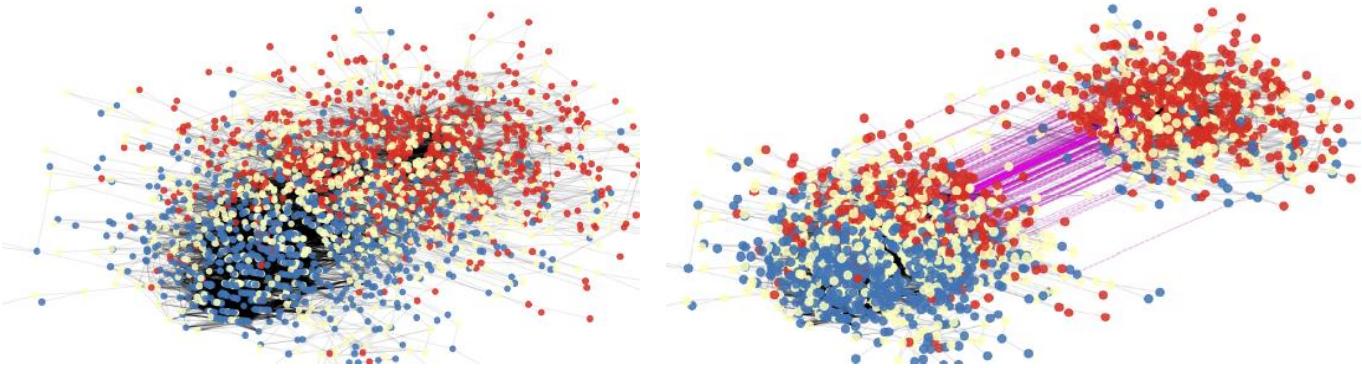
Homophily, what is it?

- Ages of pairs of friends in high school



Networks by Mark Newman

- American National Election Study data 2016



DETECTING OPINION-BASED GROUPS AND POLARISATION IN SURVEY-BASED ATTITUDE NETWORKS AND ESTIMATING QUESTION RELEVANCE

Homophily, what is it?

- Very popular idea

SPECIAL ARTICLE

The Spread of Obesity in a Large Social Network over 32 Years

Nicholas A. Christakis, M.D., Ph.D., M.P.H., and James H. Fowler, Ph.D.

- Leads to articles like this….

SECTIONS The New York Times Magazine SUBSCRIBE NOW LOG IN

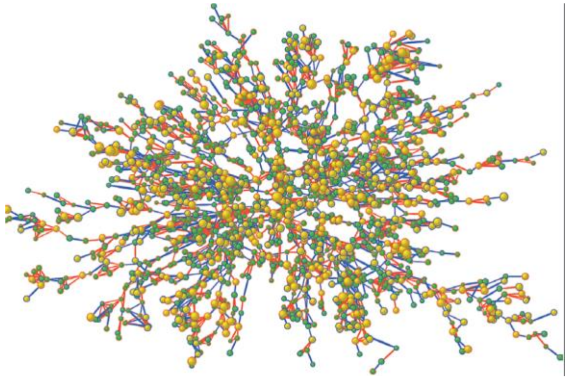
Magazine

Are Your Friends Making You Fat?

By CLIVE THOMPSON SEPT. 10, 2009

Homophily, what is it?

- Tracked peoples BMI and social contact over time
- How likely are we to be connected to other people with similar BMI?
- Made casual claims that it ‘spreads’

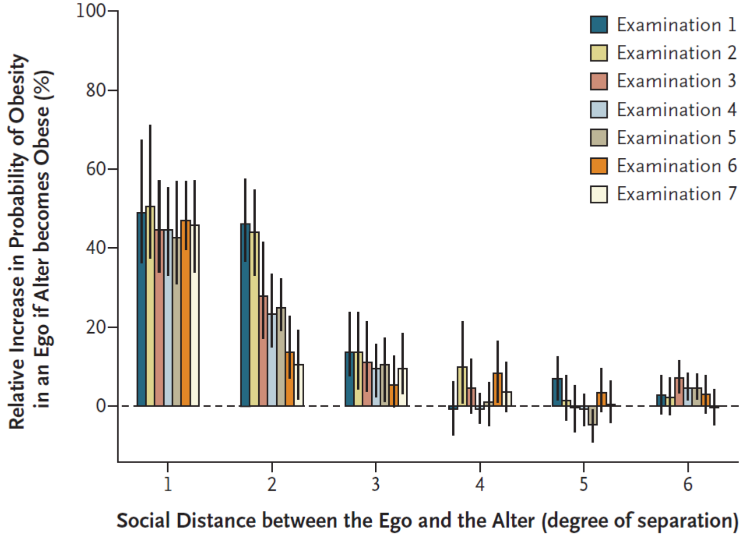


Is obesity contagious? Social networks vs. environmental factors in the obesity epidemic.

Cohen-Cole E[1], Fletcher JM.

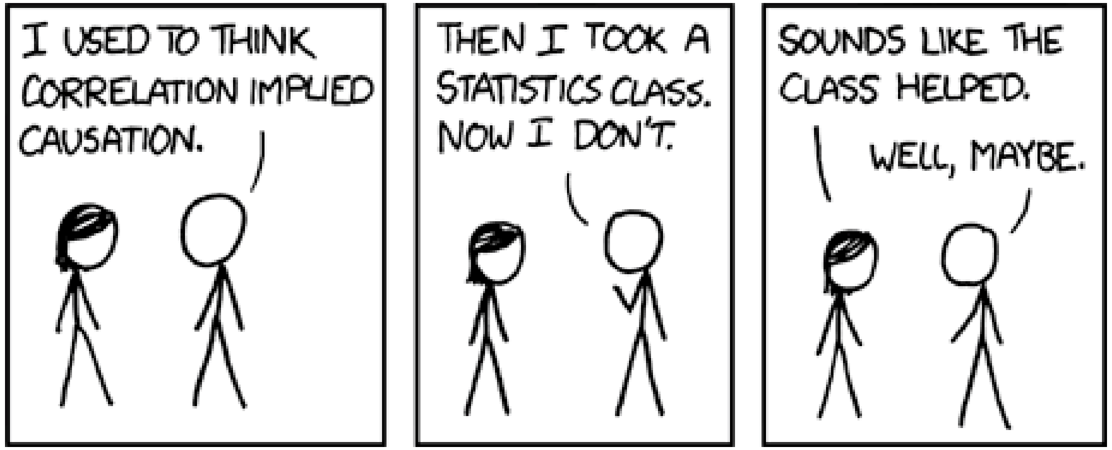
- Problem with confounders
  - Environmental factors
  - Geography
  - Age
  - Etc.

A



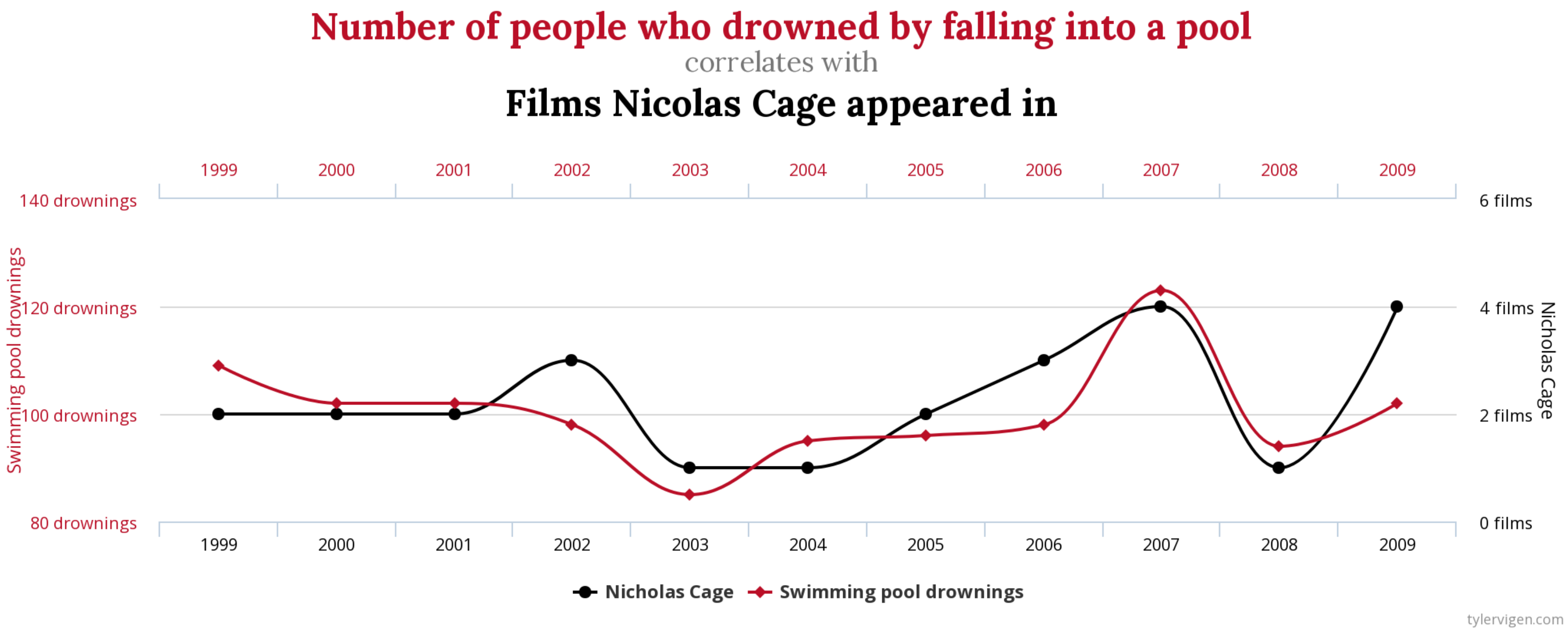
Homophily, what is it?

- Correlation does not imply causation



- Spurious correlations is a great website

Homophily, what is it?



- Ideally, in this setting, how would you prove this is ‘contagious’? (what the gold standard)?

Homophily, what is it?

- So, how can you differentiae between social and environment effects?

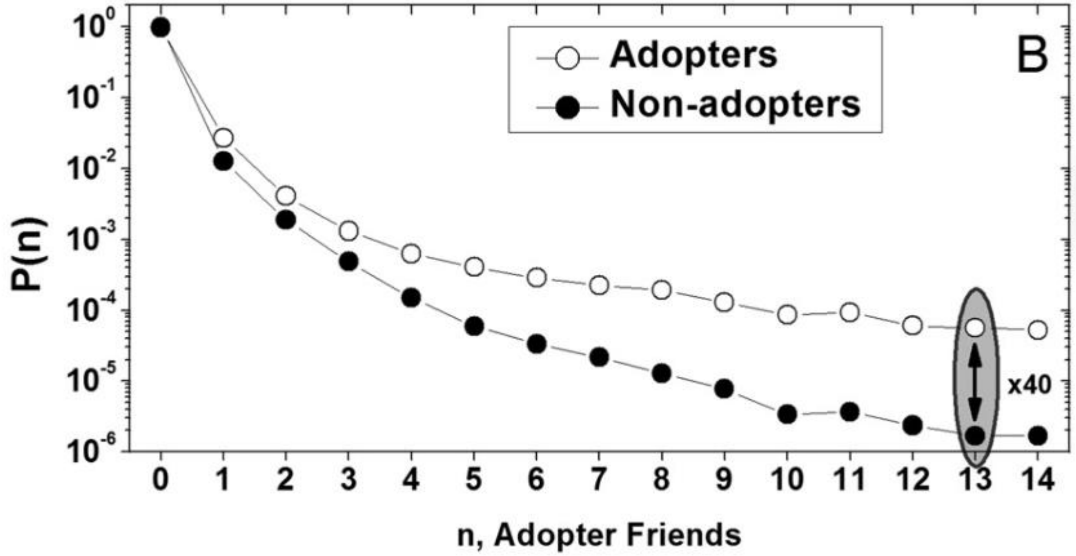
Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks

Sinan Aral, Lev Muchnik, and Arun Sundararajan

- Partnered with Yahoo
  - 30 million users Yahoo chat user
  - Over 6 months
  - Adoption of a new product
  - Extensive demographics information

Homophily, what is it?

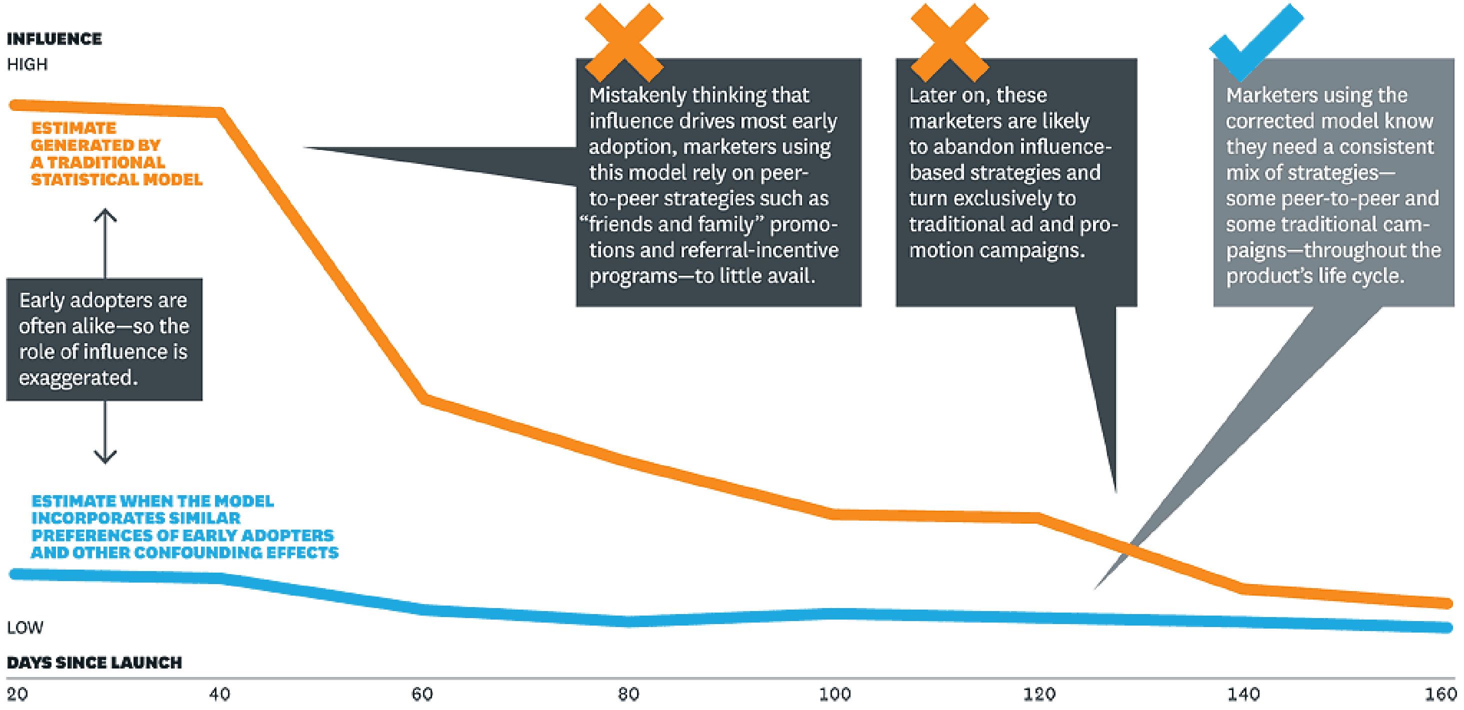
- How likely are you to about adopt given adopter neighbours



- Instead of using the raw network data
  - Dynamic match sample estimation

Homophily, what is it?

- Self sorting of people connecting via there interests



- Is this going to be a problem for us using sentiment?

Homophily, what is it?

- Actually, why do we care about doing this properly?

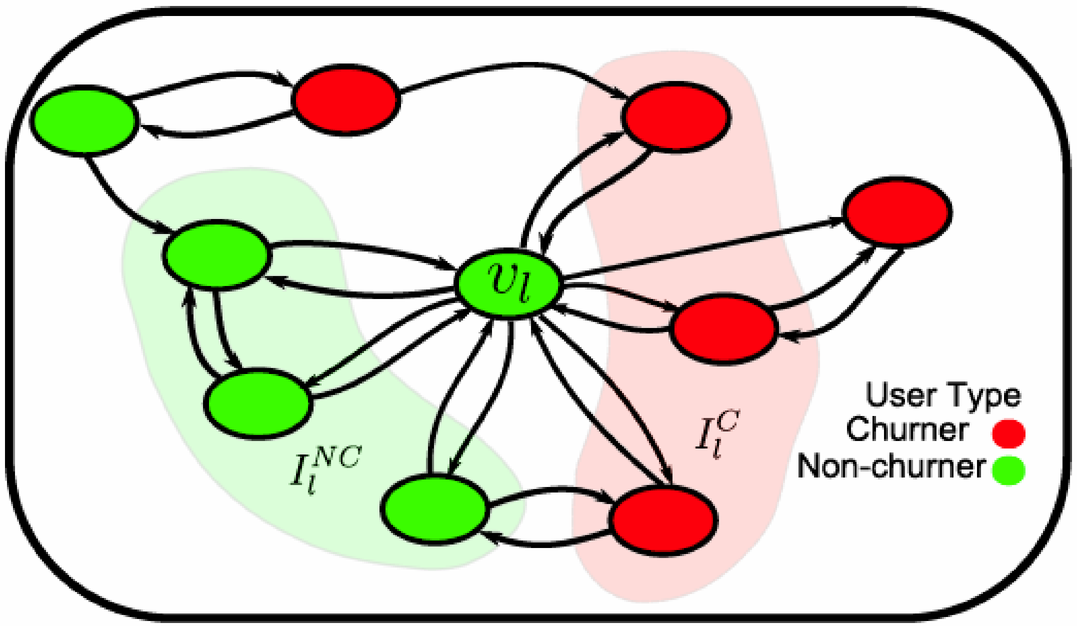
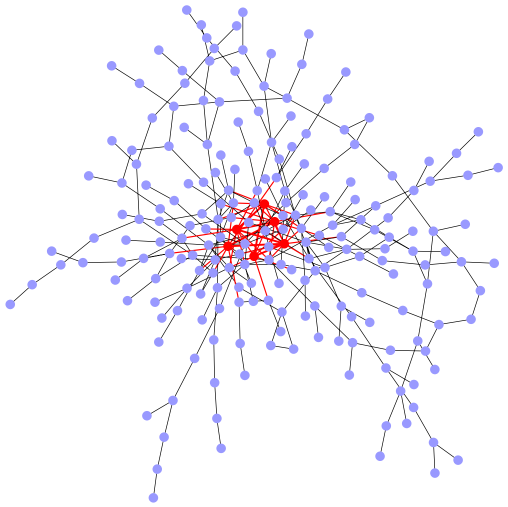
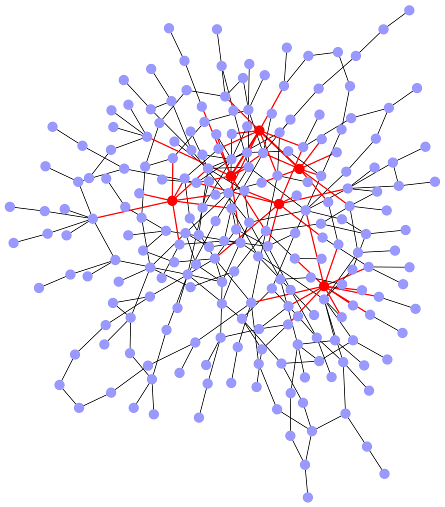


Figure 4.7: Schematic of how influence scores were calculated for each user on the network. For a non-churner node v_l, we find the total influence from churners (I_l^C) and non-churners (I_l^{NC}) as the sum of the links between v_l and their neighbours of both node types.

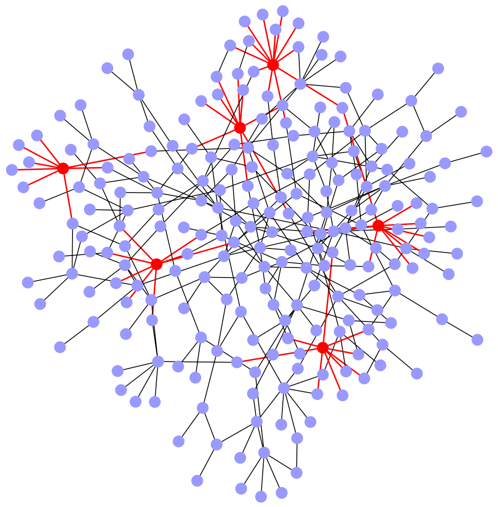
Assortativity – the other type of homophily



Assortative:
hubs show a tendency to link to each other.



Neutral:
nodes connect to each other with the expected random probabilities.



Disassortative:
Hubs tend to avoid linking to each other.

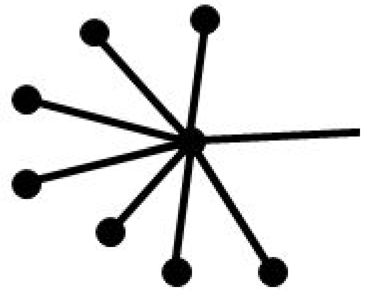
E-R Model

Configuration Model

Assorativity

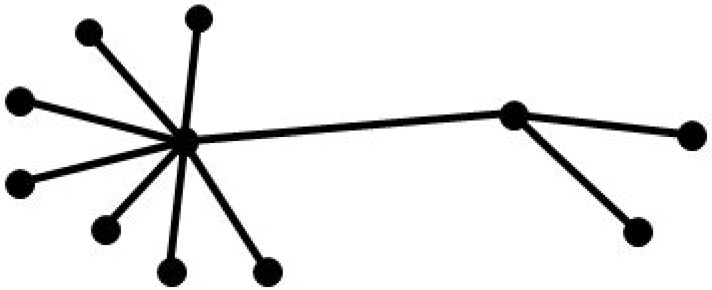
- How to calculate it?
- Pearson correlation coefficient
- But on the edges

$$q_k = \frac{(k+1)p_{k+1}}{\sum_{j\geq 1} j\, p_j}$$

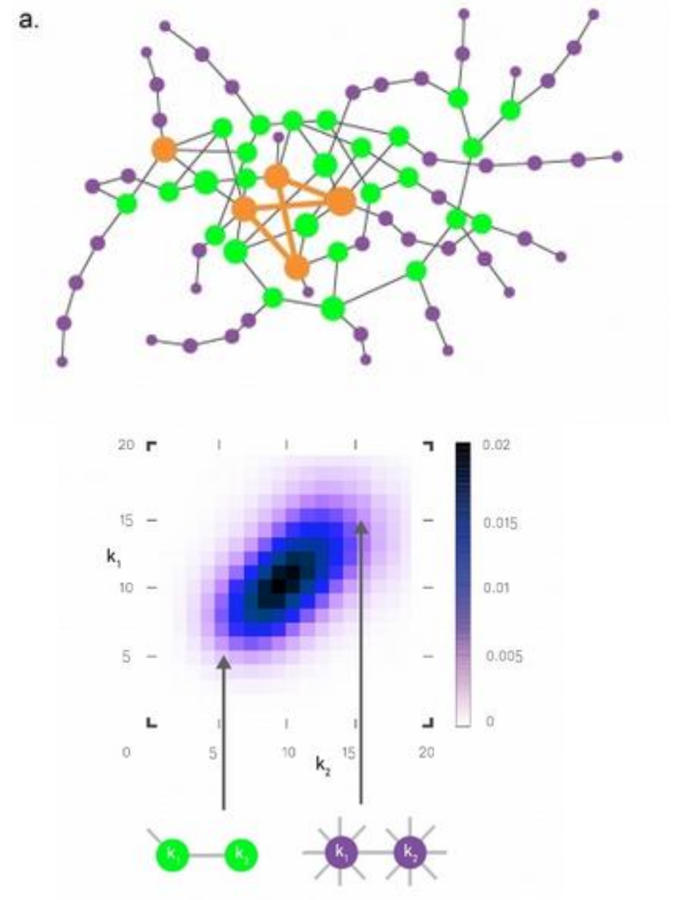


$$r = \frac{\sum_{jk} j\, k(e_{jk} - q_j q_k)}{\sigma_q^2}$$

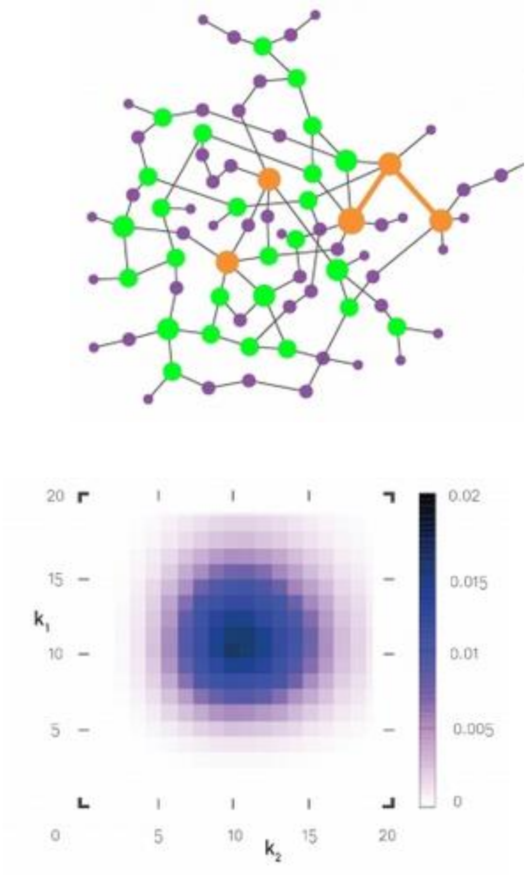
e_{jk} = Joint distibution



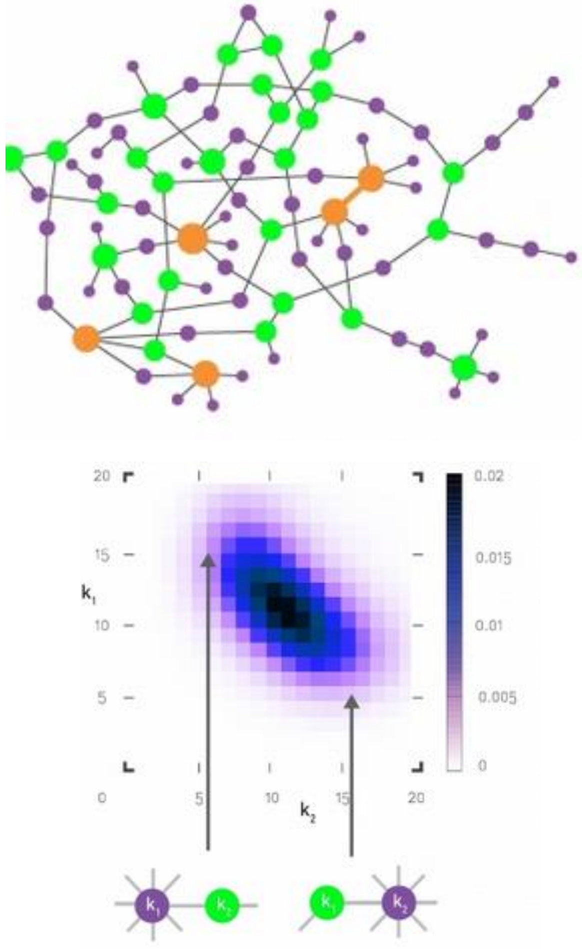
Assorativity



Assortative



Neutral



Disassortative:

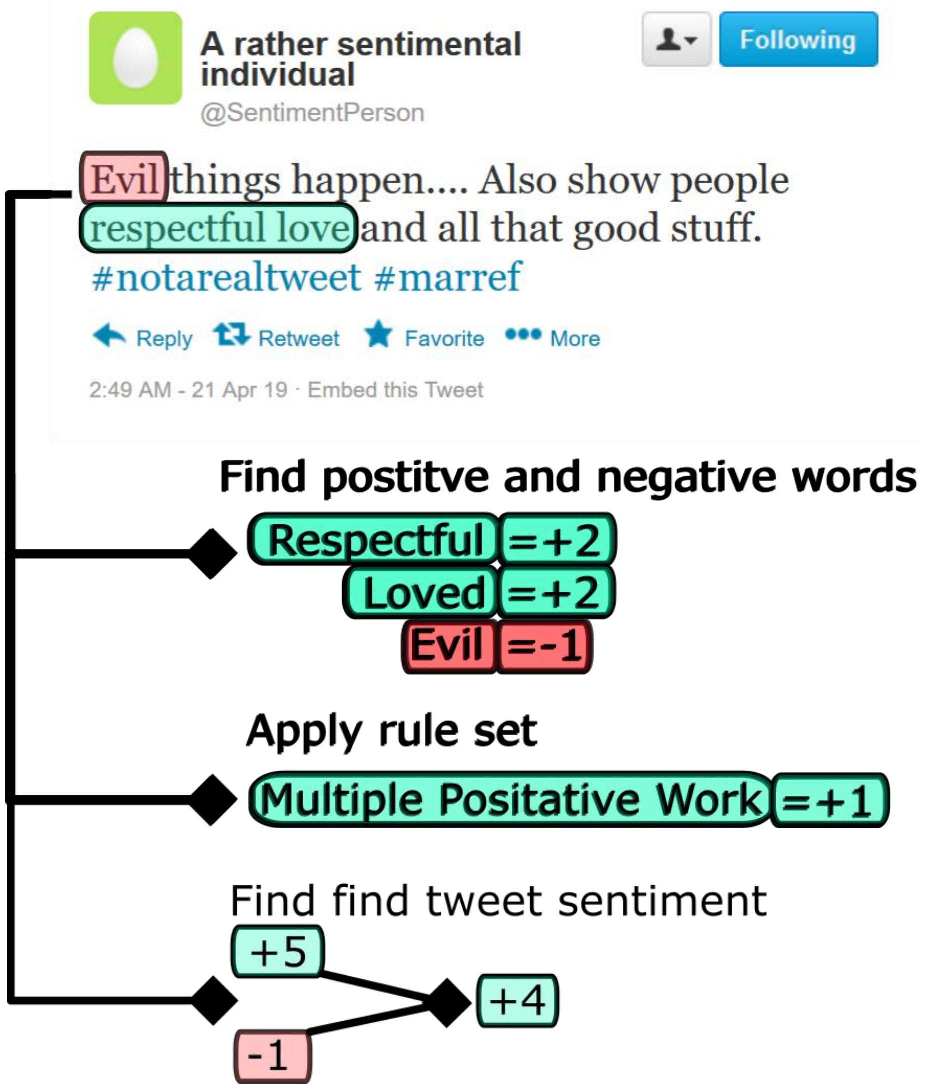
TWEET META DATA

Sentiment

http://sentistrength.wlv.ac.uk/ [1]

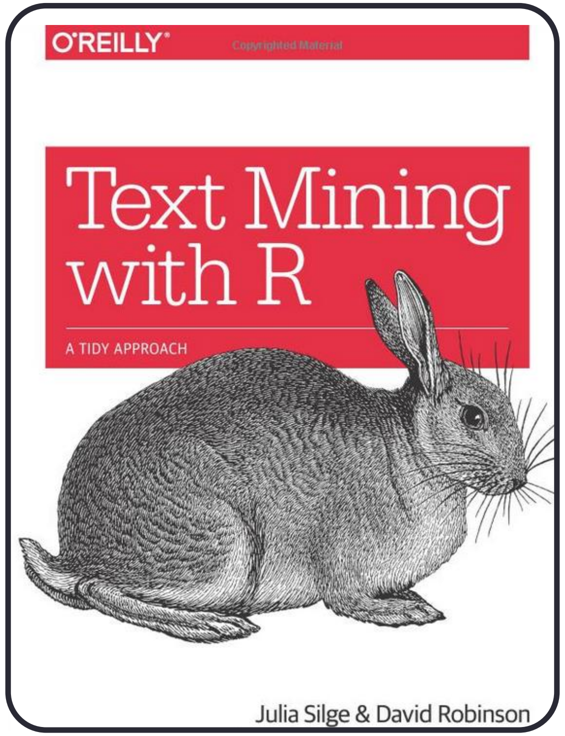
Tweet Sentiment

- Find sentiment score for each tweet
  - How positive or negative the language is
  - SentiStrength[1]
- Calculating this for every tweet
  - Fine the positive and negative score
  - Take the difference to find a single sentiment score



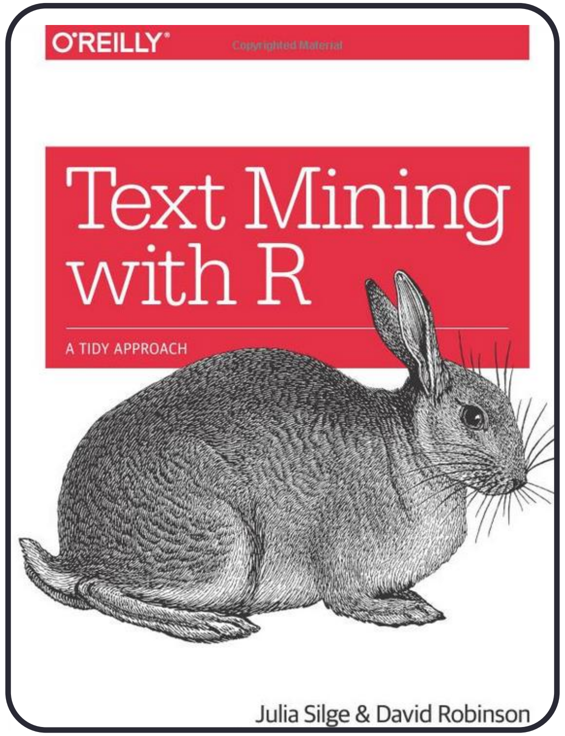
Tweet Sentiment

- Alternatives to SentiStrength



Tweet Sentiment

- Alternatives to SentiStrength



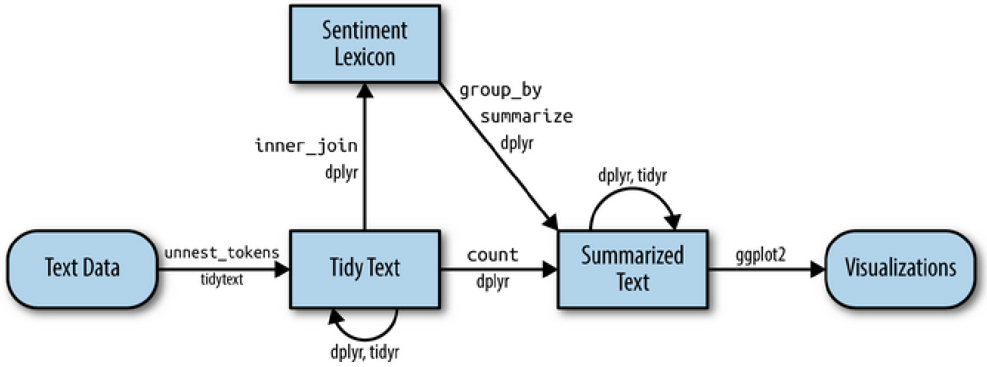


Figure 2.1: A flowchart of a typical text analysis that uses tidytext for sentiment analysis. This chapter shows how to implement sentiment analysis using tidy data principles.

Tweet Sentiment

- Alternatives to SentiStrength

2.1 The `sentiments` datasets

As discussed above, there are a variety of methods and dictionaries that exist for evaluating the opinion or emotion in text. The tidytext package provides access to several sentiment lexicons. Three general-purpose lexicons are

- `AFINN` from Finn Årup Nielsen,
- `bing` from Bing Liu and collaborators, and
- `nrc` from Saif Mohammad and Peter Turney.

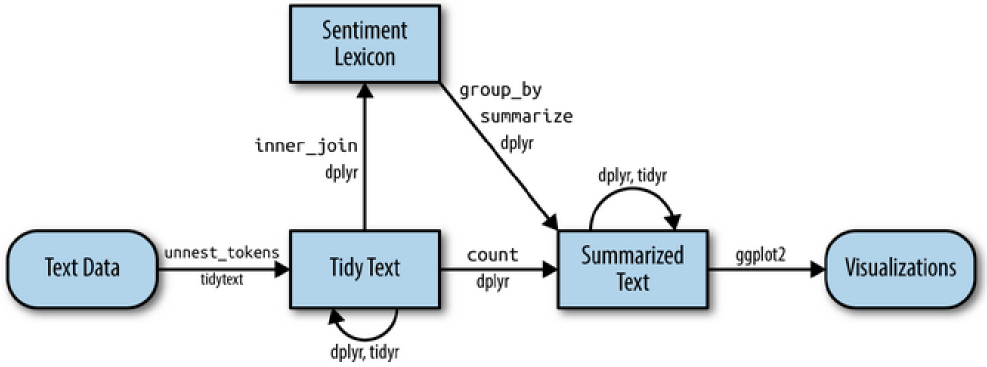
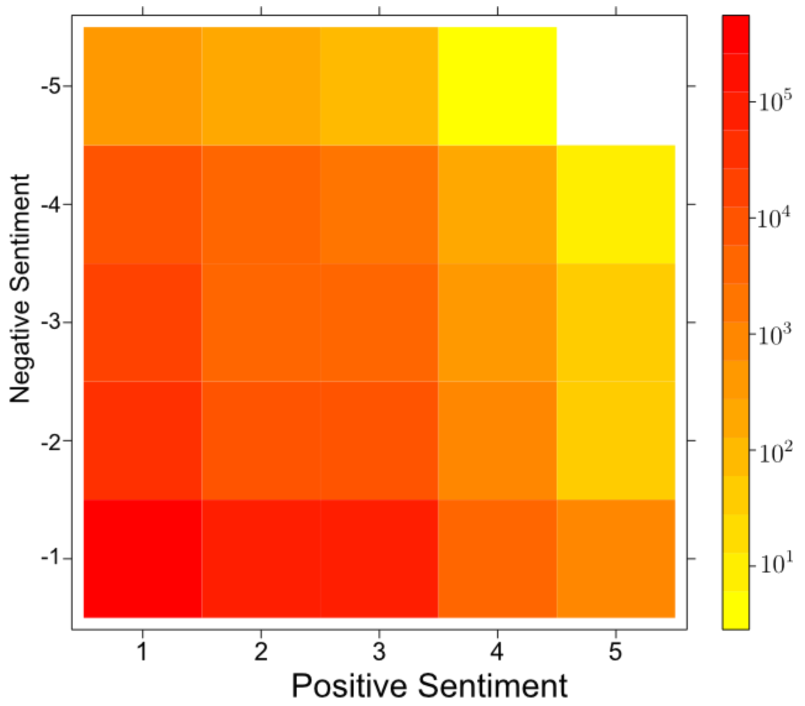


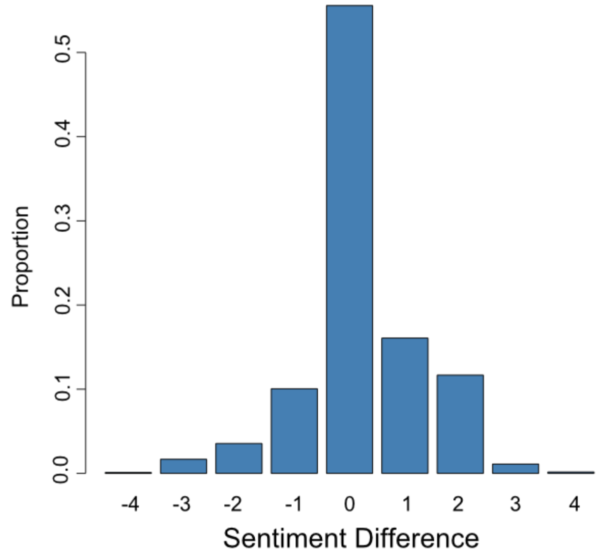
Figure 2.1: A flowchart of a typical text analysis that uses tidytext for sentiment analysis. This chapter shows how to implement sentiment analysis using tidy data principles.

Tweet Sentiment

- Distribution of scores
    - Scale from -4 to 4
    - Most tweets have 0 sentiment score (55%)
    - Vast majority had (-1,1) sentiment scores (95%)

- Sentiment is noisy
    - Look to aggregate out the noise
    - Distribution of tweets per user
        - Average 4
        - Problematic for aggregation

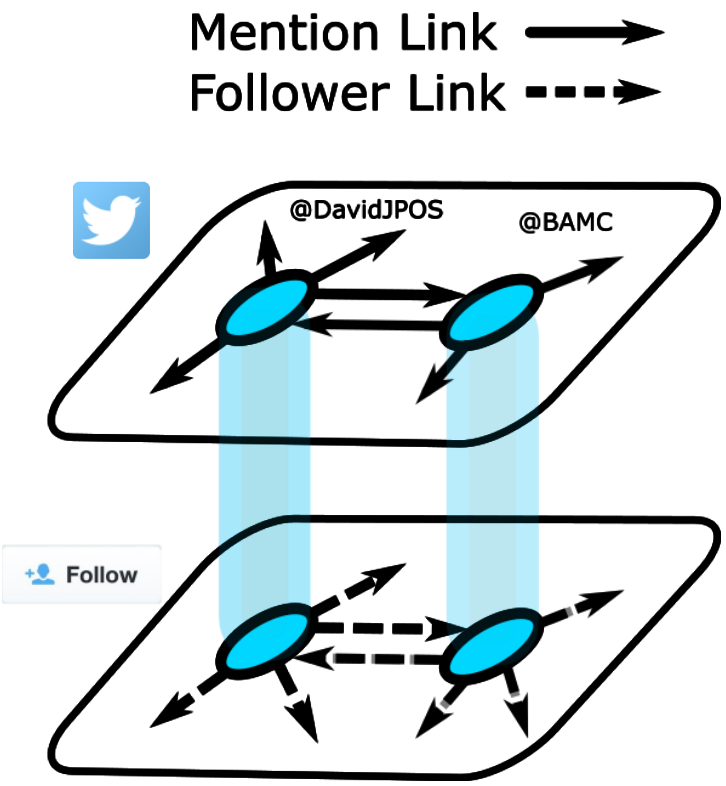




Network creation

- Generate two networks
  - Mention – conversational
  - Follower – structural

  - Weight mention links by sentiment

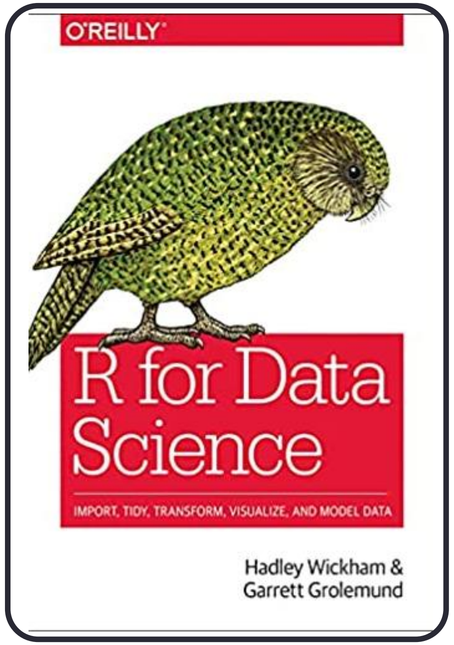


TWEET META DATA

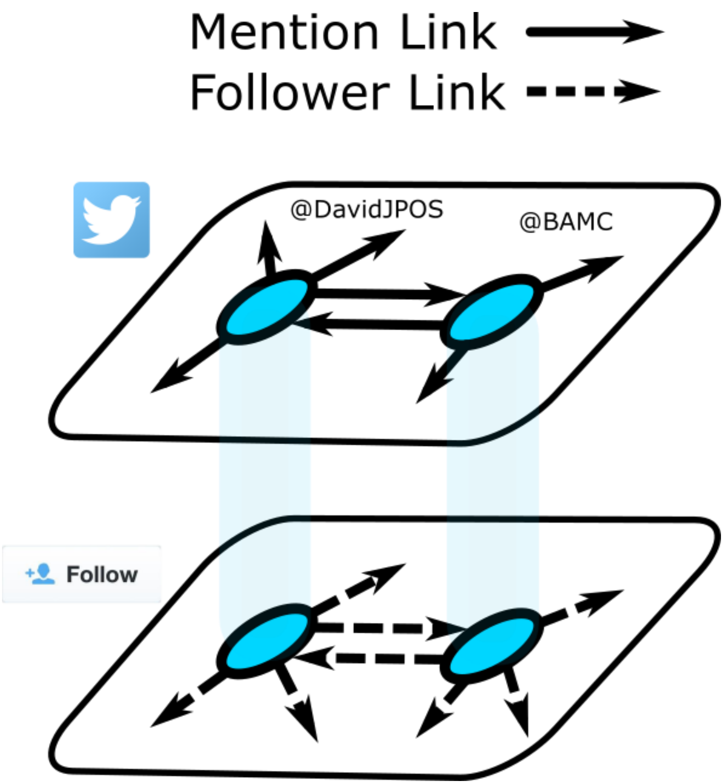
A little aside on how we extra the network from tweets

Network creation

- How does this actually work?
  - Chapter 14 in R for data science
  - stringr packages <- a little text processing



stringr

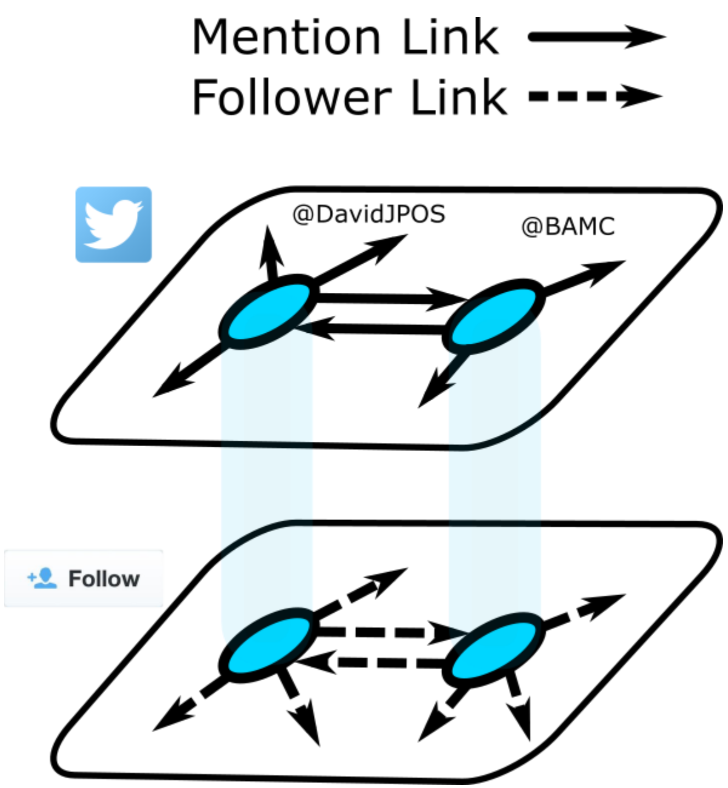


Network creation

- How does this actually work?

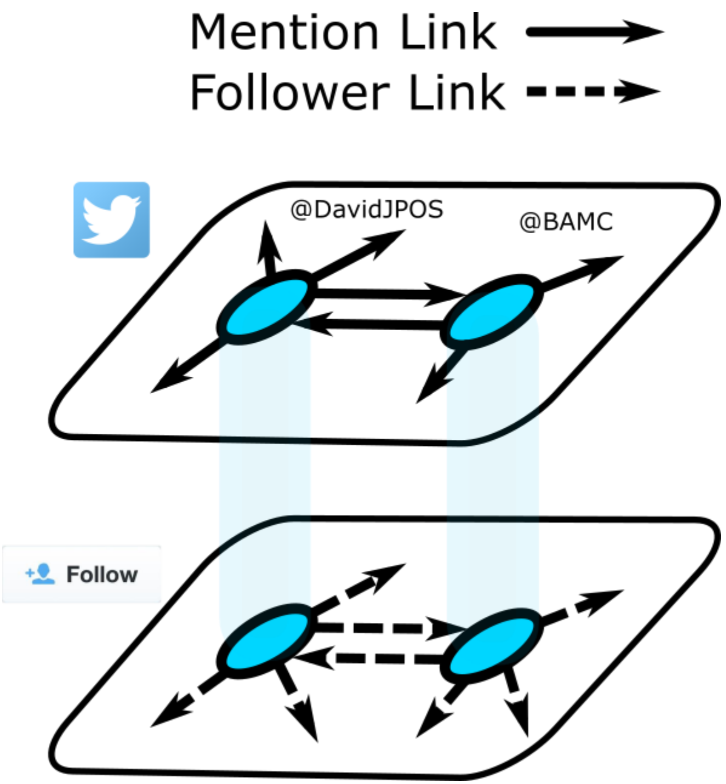
- Regular expressions to extract these
  - a sequence of characters that specifies a search pattern

- https://regexr.com/

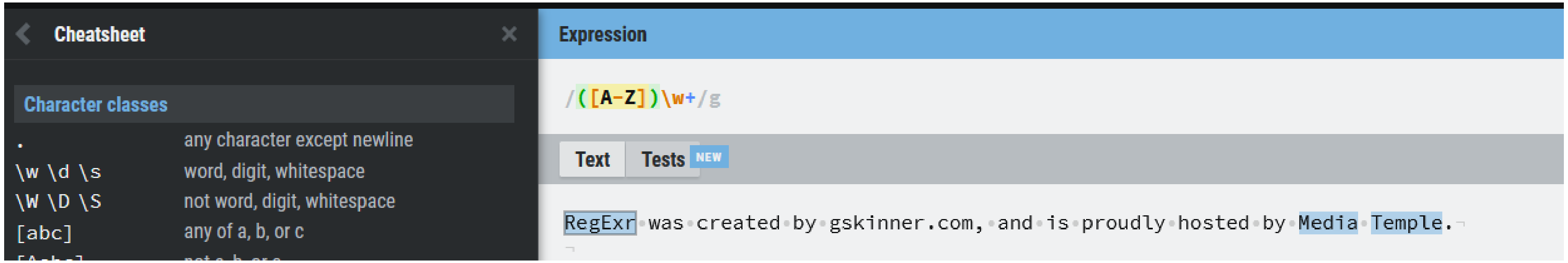


Network creation

- How does this actually work?

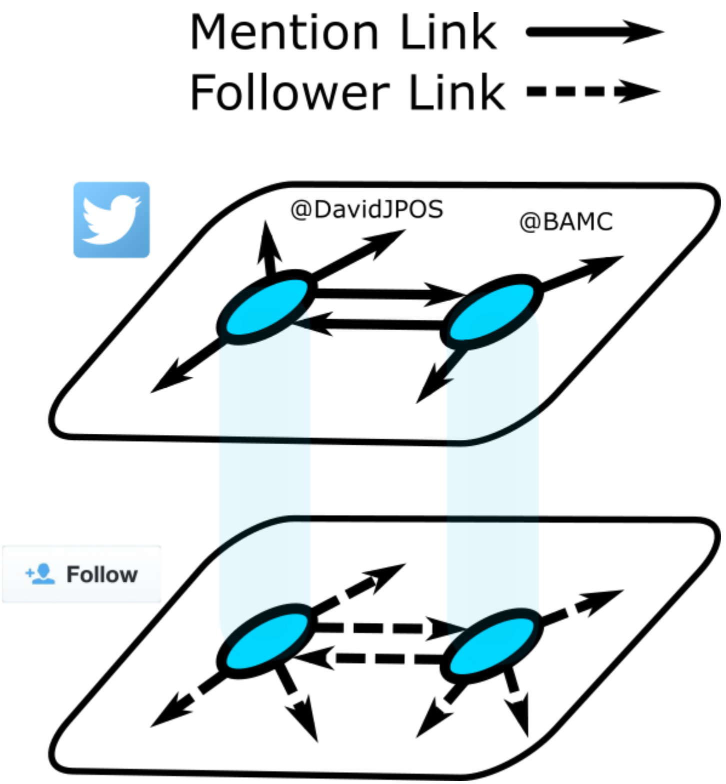


- Regular expressions to extract these
- https://regexr.com/



Network creation

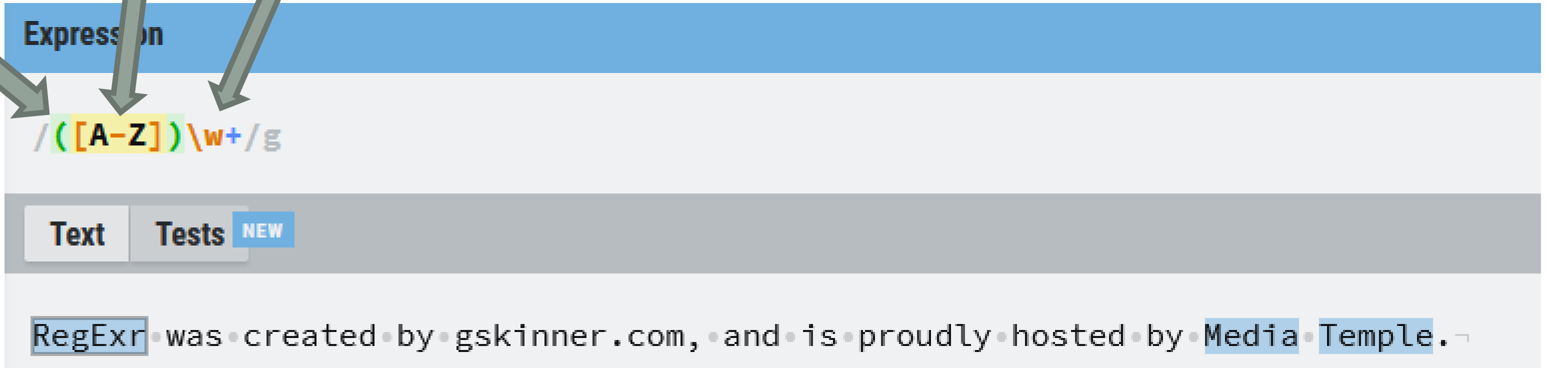
- How does this actually work?



Any of A-Z works

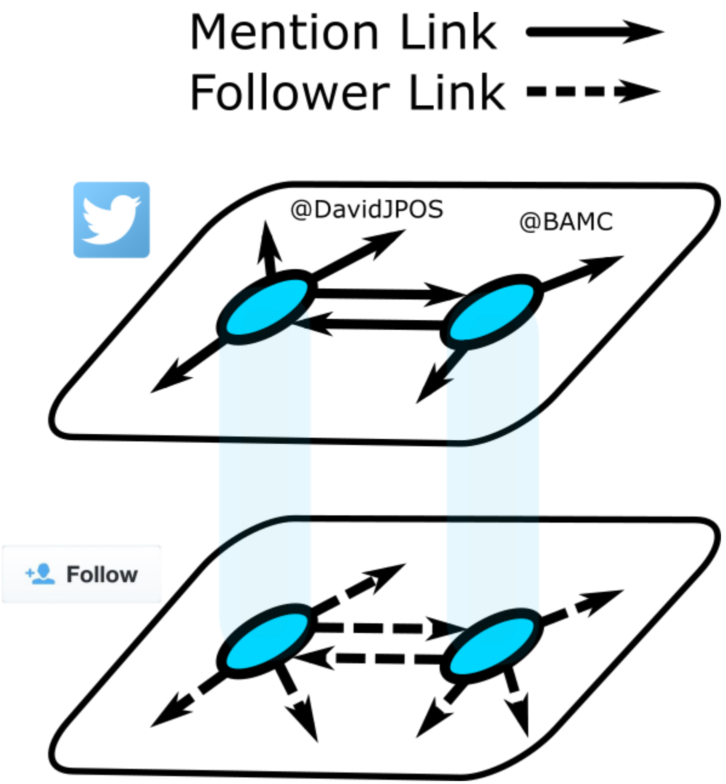
Defines 'capture group'

And rest of the word

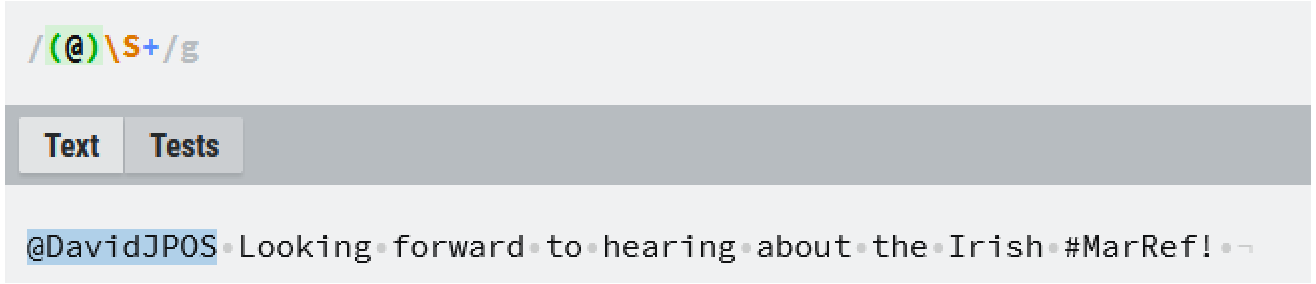


Network creation

- How does this actually work?

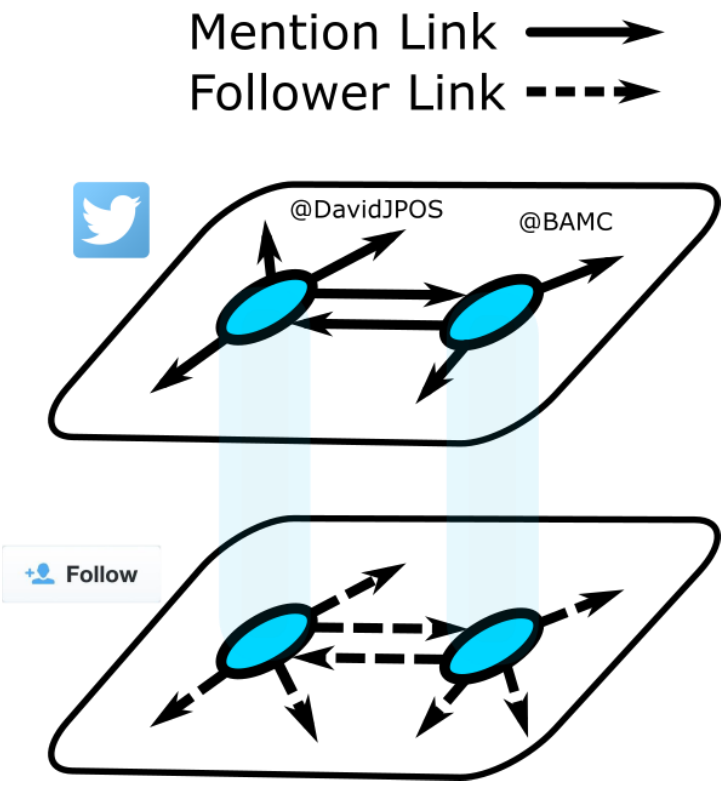


- Regular expressions to extract these

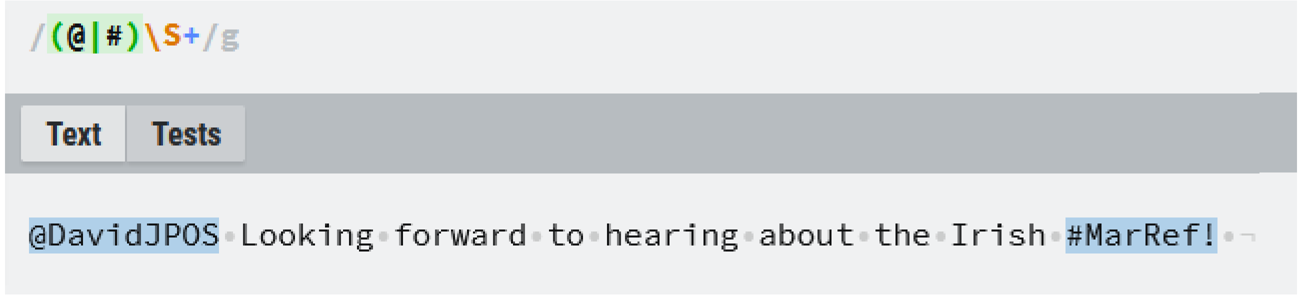


Network creation

- How does this actually work?



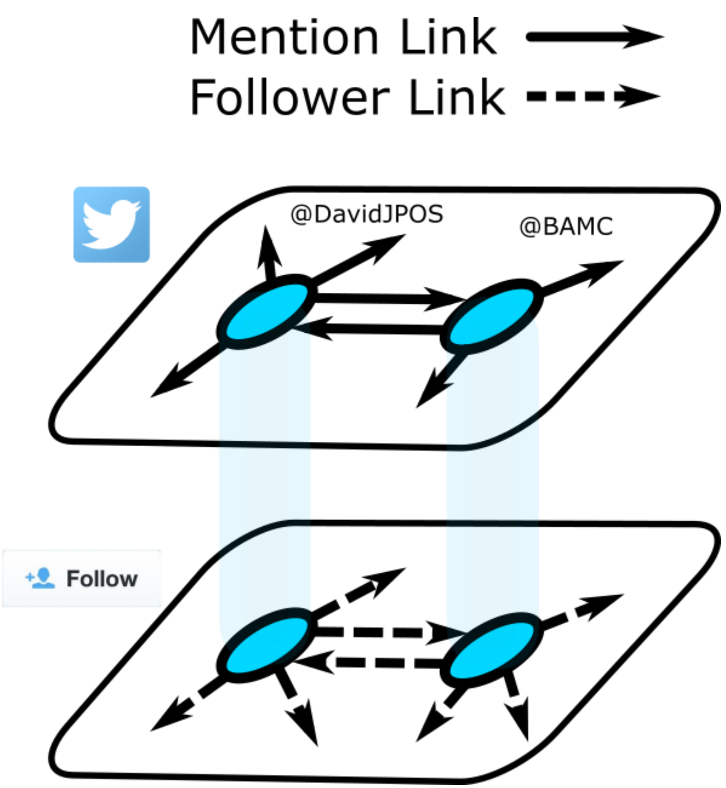
- Regular expressions to extract these



Network creation

- How does this actually work?

- Regular expressions to extract these



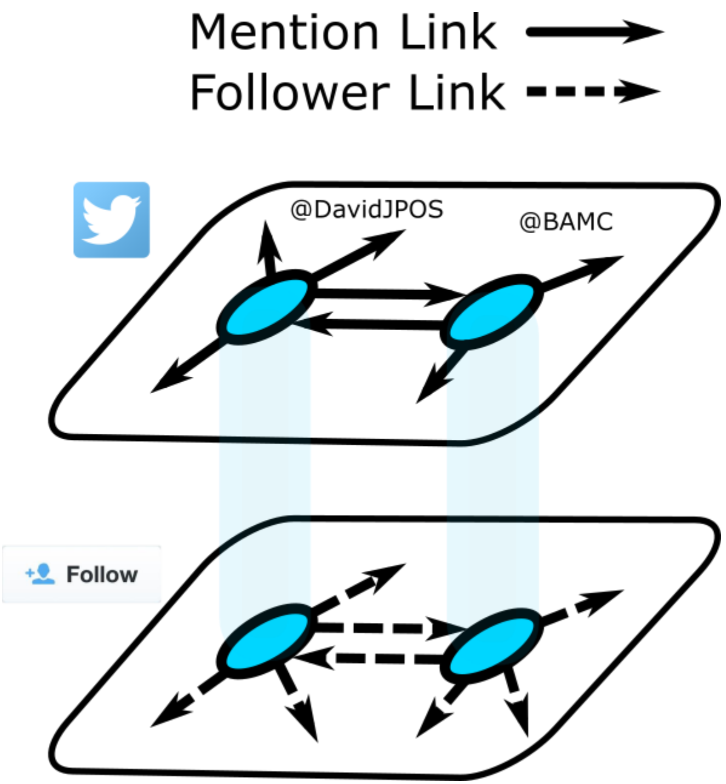

```
> text <- '@DavidJPOS Looking forward to hearing about the Irish #MarRef!'
> stringr::str_extract_all(string = text, pattern = '(@|#)\\S+')
[[1]]
[1] "@DavidJPOS" "#MarRef!"

>
```

Network creation

- How does this actually work?

- Regular expressions to extract these

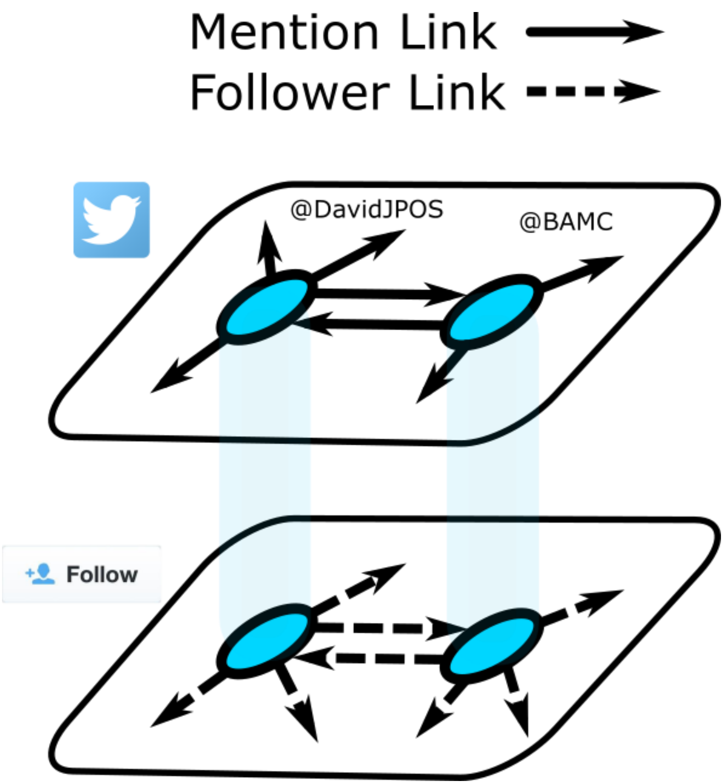



```
> text <- 'RT @someone: @DavidJPOS Looking forward to hearing about the Irish #MarRef!'
> text %>%
+   stringr::str_extract_all(pattern = '(@|#)\\S+')
[[1]]
[1] "@someone:"  "@DavidJPOS" "#MarRef!"
```

Network creation

- How does this actually work?

- Regular expressions to extract these

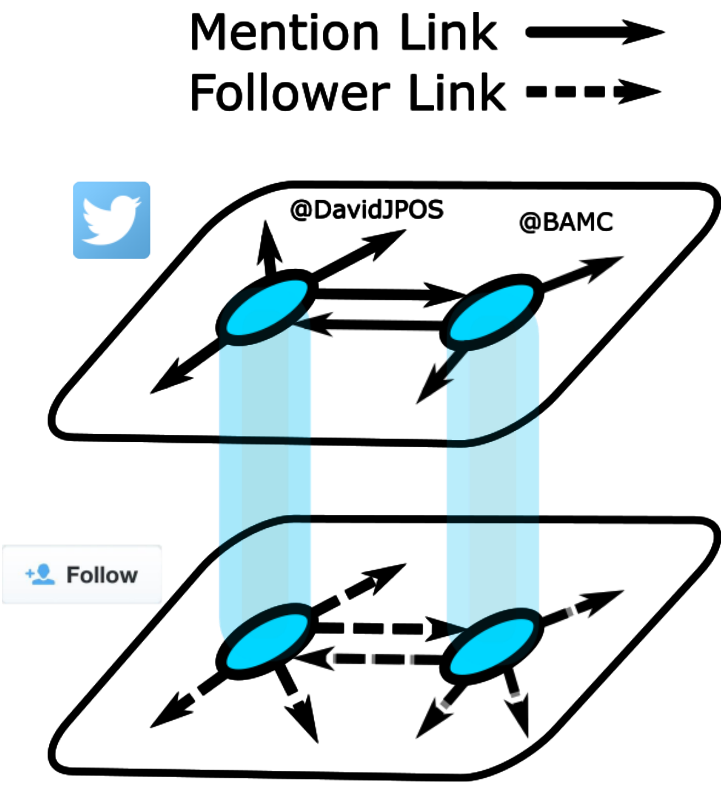



```
> text <- 'RT @someone: @DavidJPOS Looking forward to hearing about the Irish #MarRef!'
> text %>%
+   stringr::str_remove(pattern = 'RT @\\S+') %>%
+   stringr::str_extract_all(pattern = '(@|#)\\S+')
[[1]]
[1] "@DavidJPOS" "#MarRef!"
```

Network creation

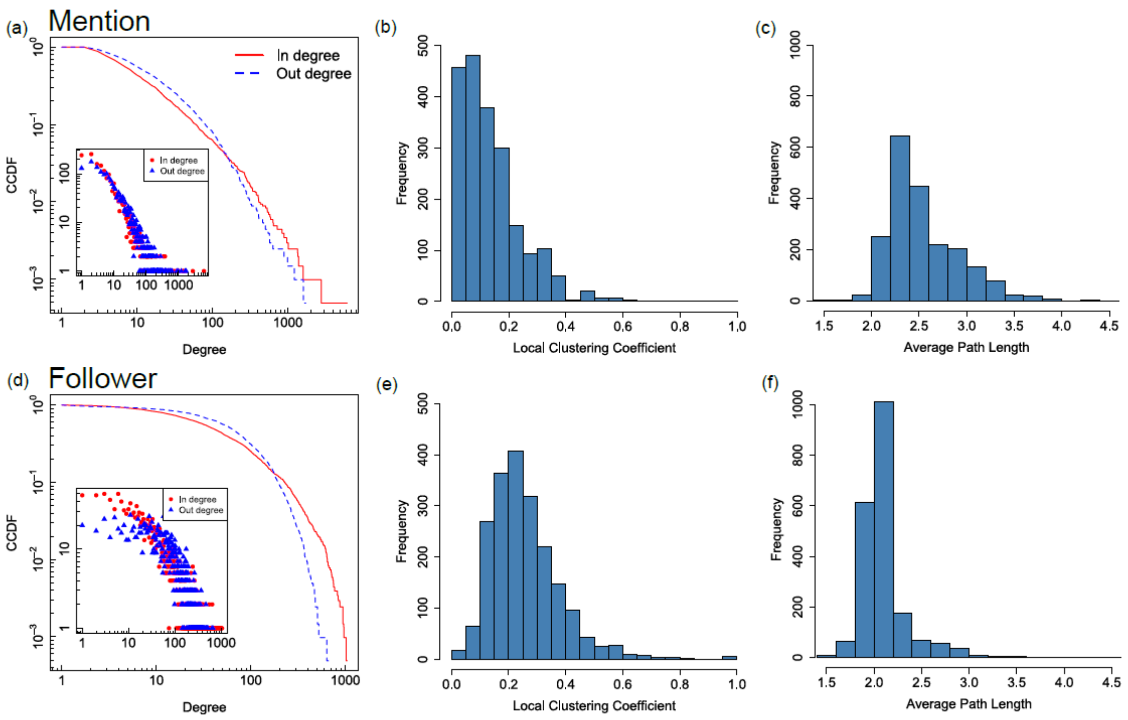
- Generate two networks
  - Mention – conversational
  - Follower – structural

  - Weight mention links by sentiment



Network creation

	Mention network		Follower network	
	Full	Reciprocal	Full	Reciprocal
Nodes	40,812	2,047	36,674	2,047
Links	227,203	69,022	3,309,687	173,137
Reciprocal links	23,713	22,218	1,398,236	85,986
Avg. out degree	9	34	90	85
Transitivity	0.02	0.13	0.09	0.28



Another little break to play with R

- Create the empirical network from data

  - 2_descriptive.r

NETWORK GENERATION

How do we create useful simulated networks?

Erdős-Rényi networks

- To construct a random network we follow these steps:
  - Start with *N* isolated nodes.
  - Select a node pair, create an edge with prob p
  - Repeat for each of the *N*(*N*-1)/2 node pairs.

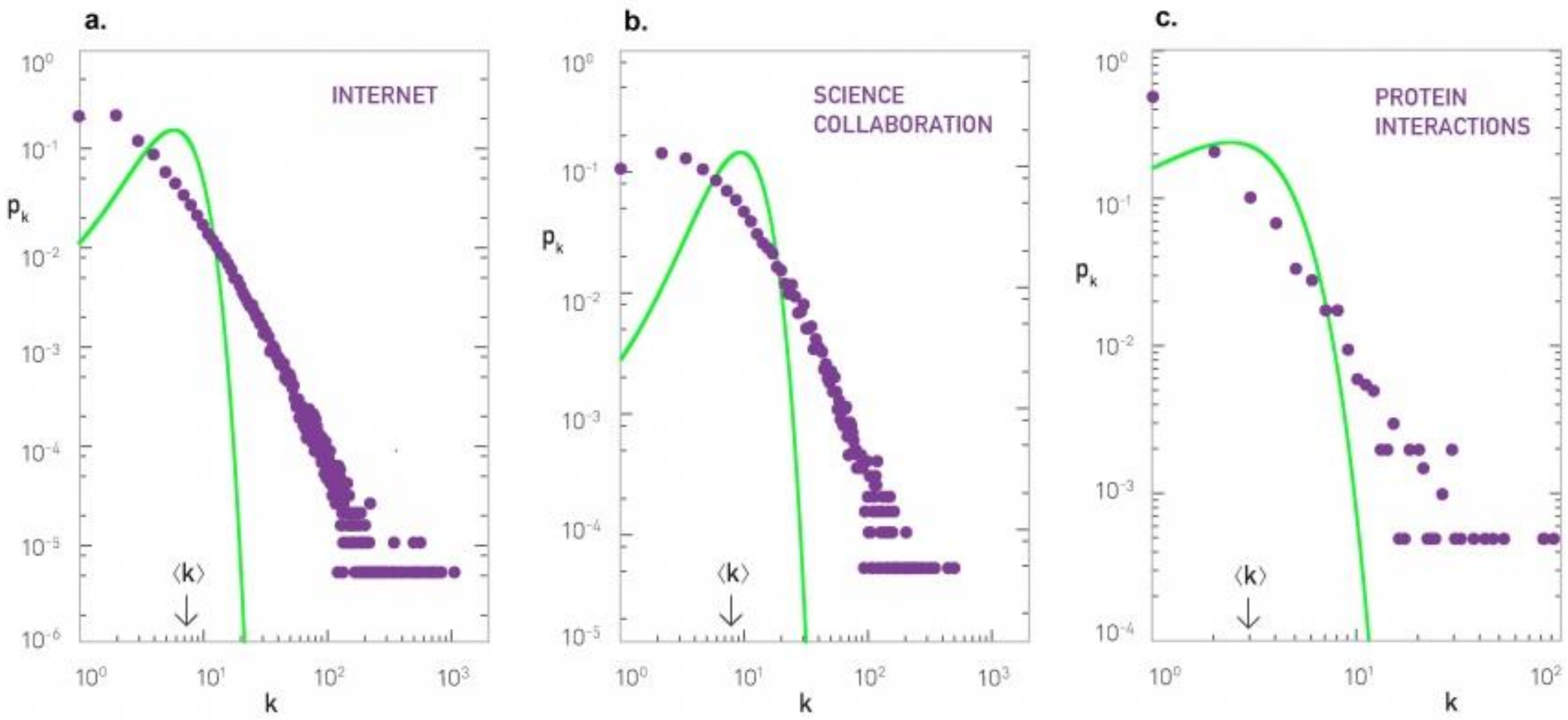
  - Node number of links follows a binomial distribution

$$p_k = \binom{N-1}{k} p^k (1-p)^{N-1-k}$$

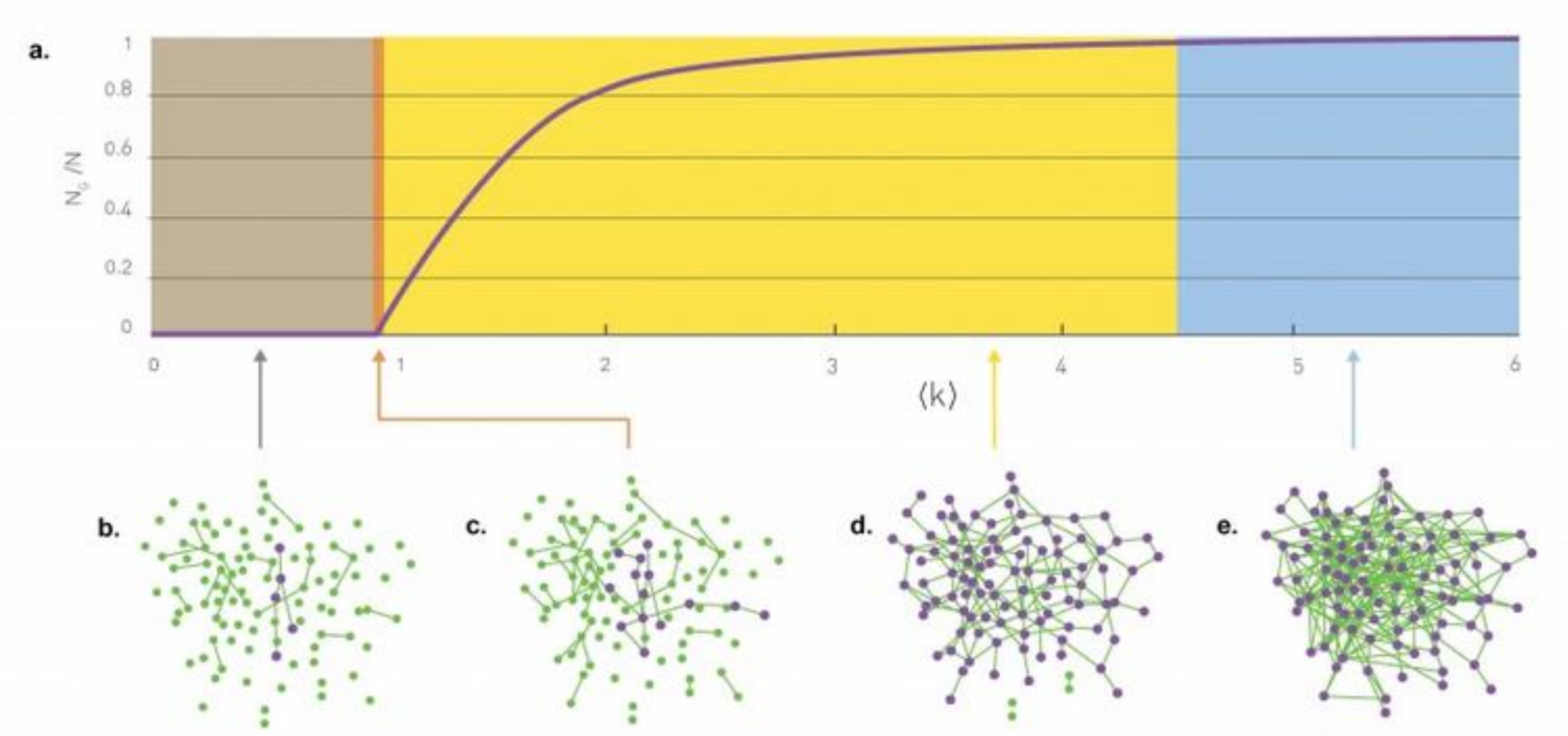
  - But as N get large a binomial follows a normal distribution

$$p_k = e^{-\lambda} \frac{\lambda^k}{k!}$$

Erdős-Rényi networks



Erdős-Rényi networks

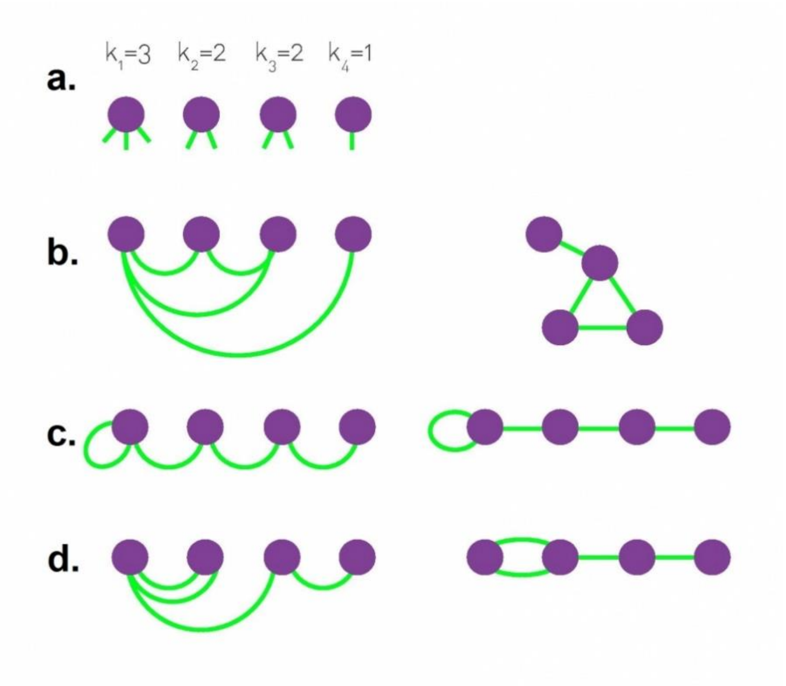


Network generation models

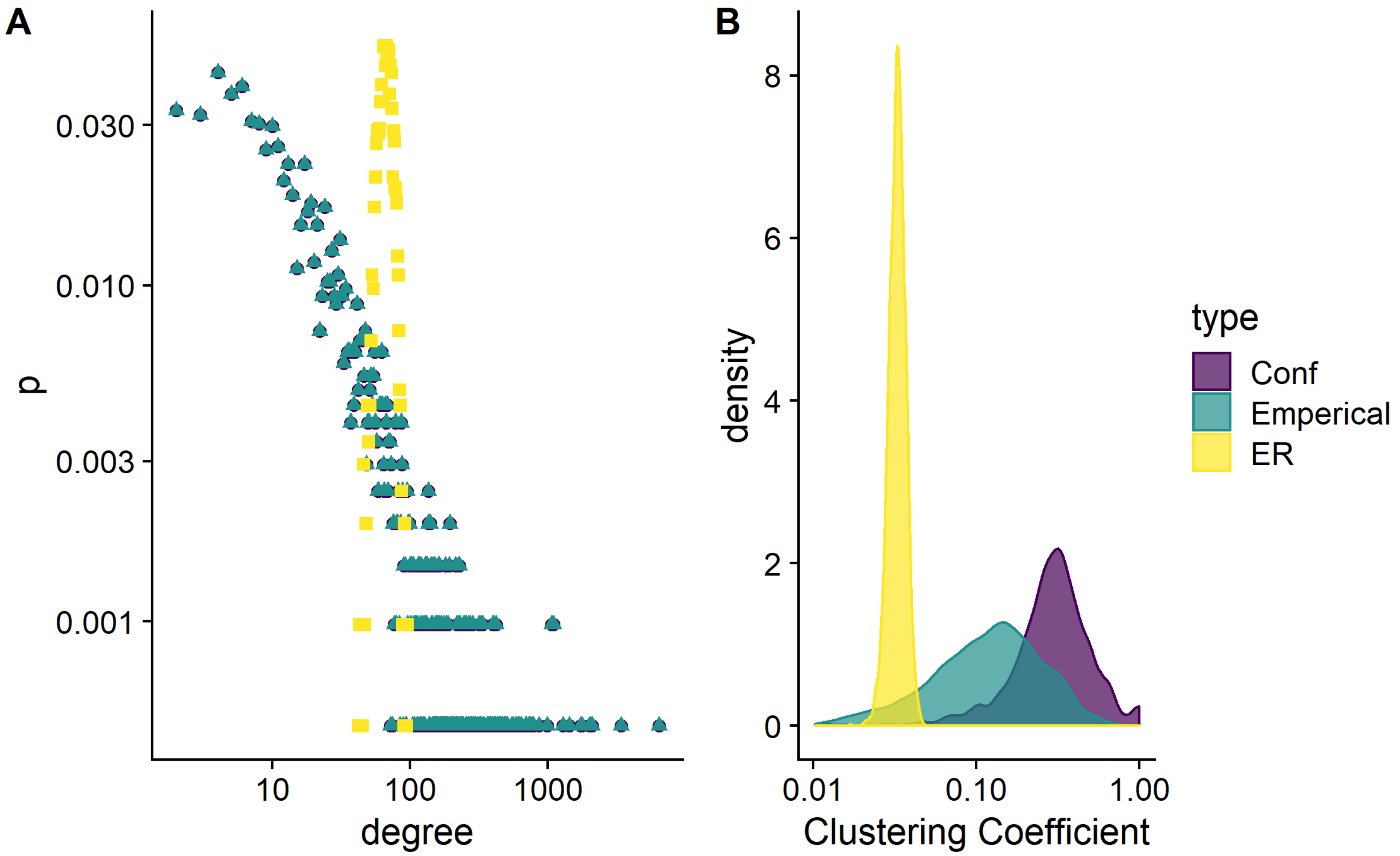
- What about a predefined degree dist?
- Configuration model
  - Defined degree sequence
  - Randomly connect ‘stubs’ together

- Can contain
  - Multi edges
  - Self loops
  - But for large N these are small

- Useful model analytically
- But how close to empirical networks is it?



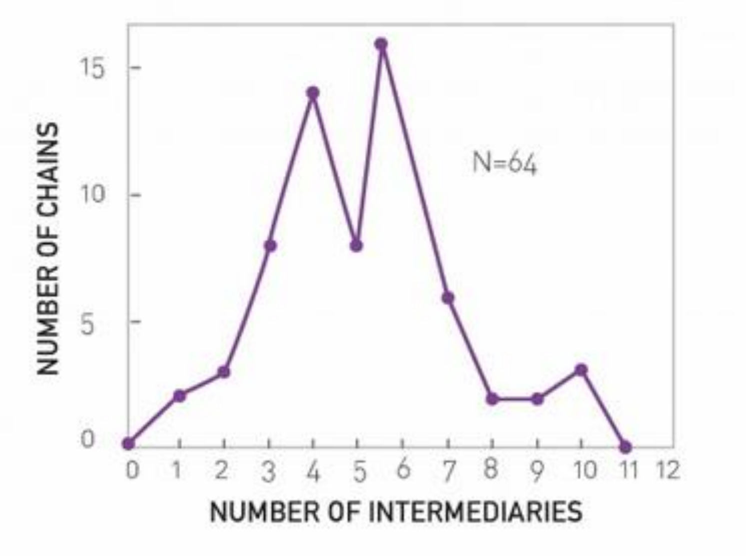
Network generation methods

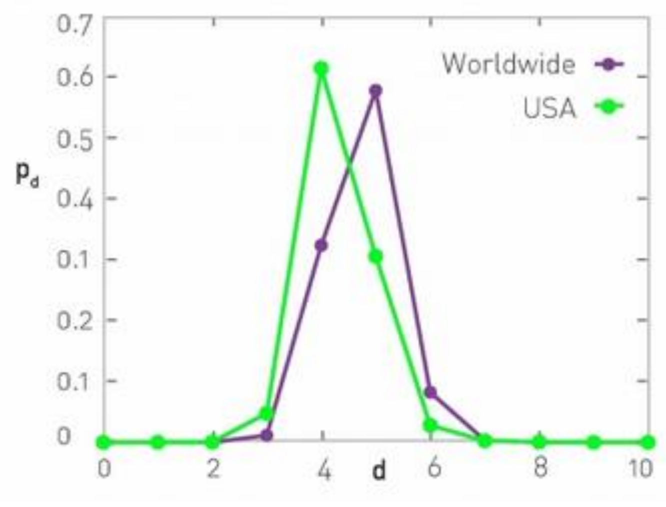


Small World phenomenon

- Six Degree of Separation
  - Stanley Milgram – 1967 letter passing
  - experiment to measure the distances in social networks
  - eventually 64 of the 296 letters made it back

  - Facebook’s social graph
  - 721 million active users
  - 68 billion symmetric friendship links
  - average distance 4.74 between the users





Small World phenomenon

- Six Degree of Separation

Network	N	L	‹k›	‹d›	d_{max}
Internet	192,244	609,066	6.34	6.98	26
WWW	325,729	1,497,134	4.60	11.27	93
Power Grid	4,941	6,594	2.67	18.99	46
Mobile-Phone Calls	36,595	91,826	2.51	11.72	39
Email	57,194	103,731	1.81	5.88	18
Science Collaboration	23,133	93,437	8.08	5.35	15
Actor Network	702,388	29,397,908	83.71	3.91	14
Citation Network	449,673	4,707,958	10.43	11.21	42
E. Coli Metabolism	1,039	5,802	5.58	2.98	8
Protein Interactions	2,018	2,930	2.90	5.61	14

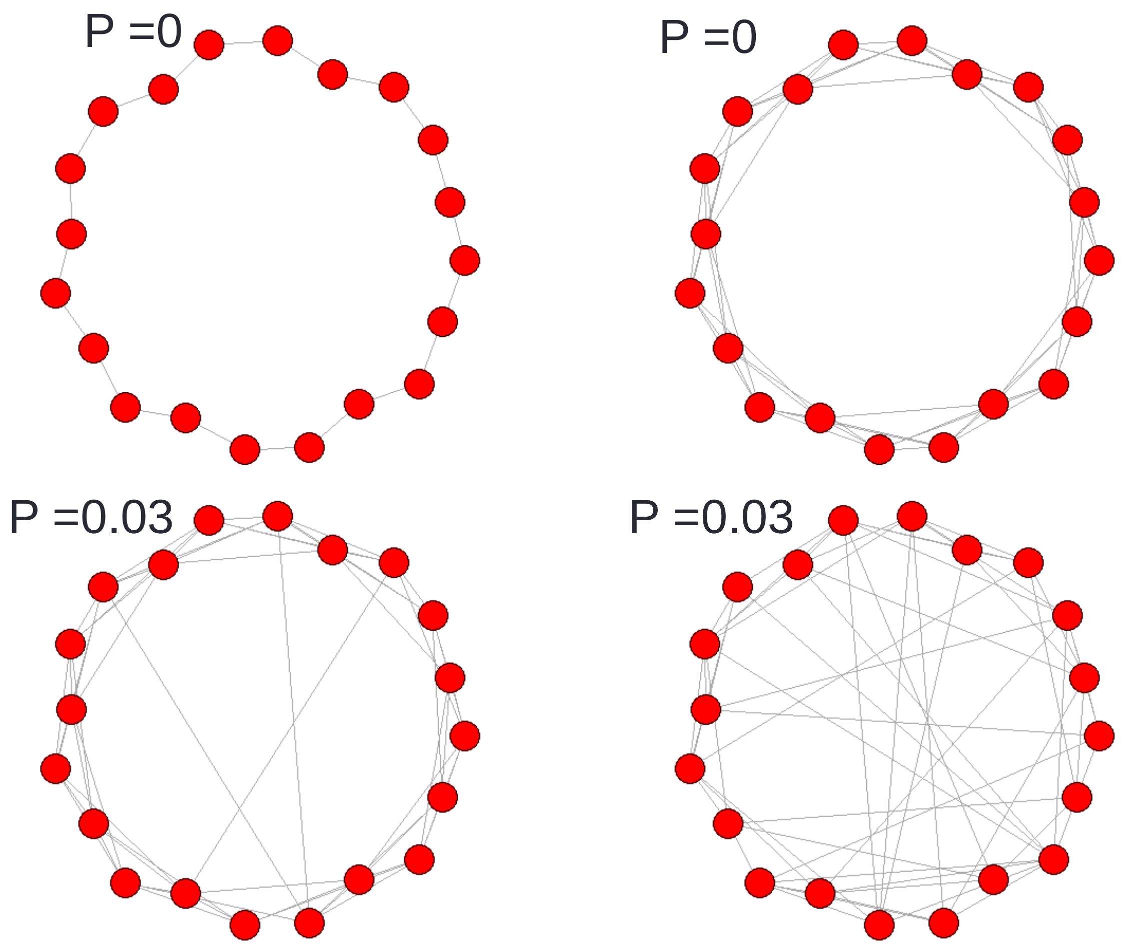
Small World phenomenon

- Six Degree of Separation

Network	N	L	‹k›	‹d›	d_{max}
Internet	192,244	609,066	6.34	6.98	26
WWW	325,729	1,497,134	4.60	11.27	93
Power Grid	4,941	6,594	2.67	18.99	46
Mobile-Phone Calls	36,595	91,826	2.51	11.72	39
Email	57,194	103,731	1.81	5.88	18
Science Collaboration	23,133	93,437	8.08	5.35	15
Actor Network	702,388	29,397,908	83.71	3.91	14
Citation Network	449,673	4,707,958	10.43	11.21	42
E. Coli Metabolism	1,039	5,802	5.58	2.98	8
Protein Interactions	2,018	2,930	2.90	5.61	14

Small world network: Watts-Strogatz network

- Building in transitivity
- p controls random rewiring
- Creates short paths through the network – hence ‘small world’
- ‘Strength of weak ties hypothesis’



Another little break to play with R

- Network generation models

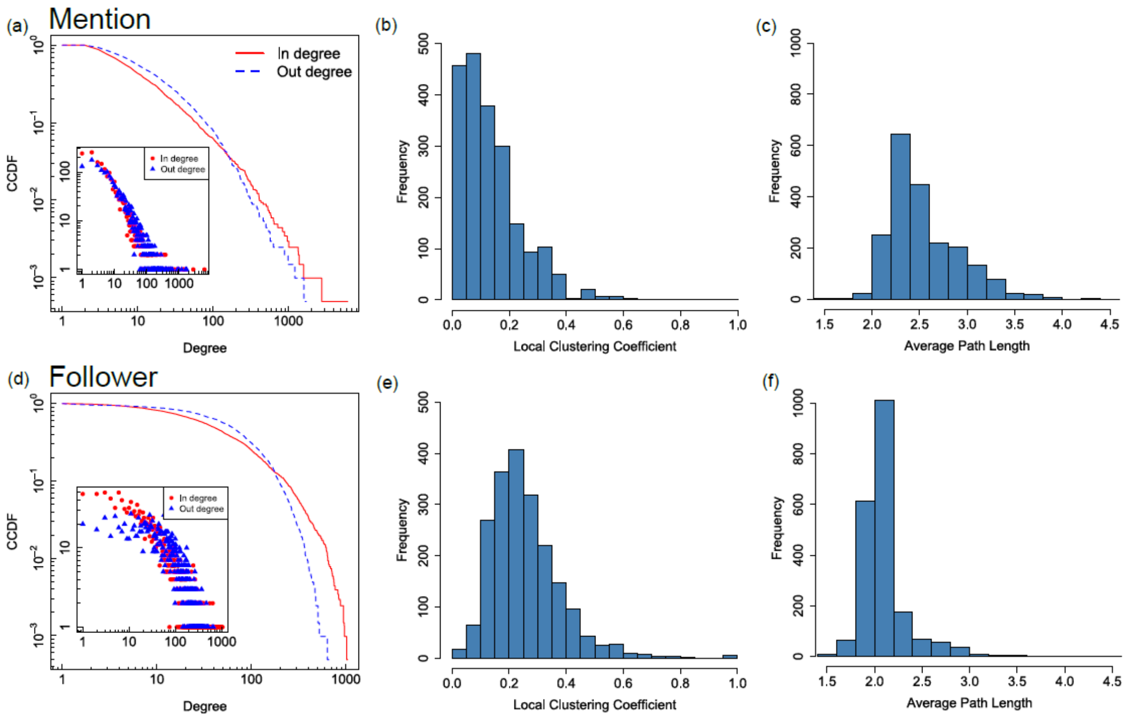
- How would we generate similar networks to empirical observed networks?
  - Statistical networks
  - Ensembles properties of these random network

  - 2_descriptive.r

RANDOMISATION TESTS FOR NETWORKS

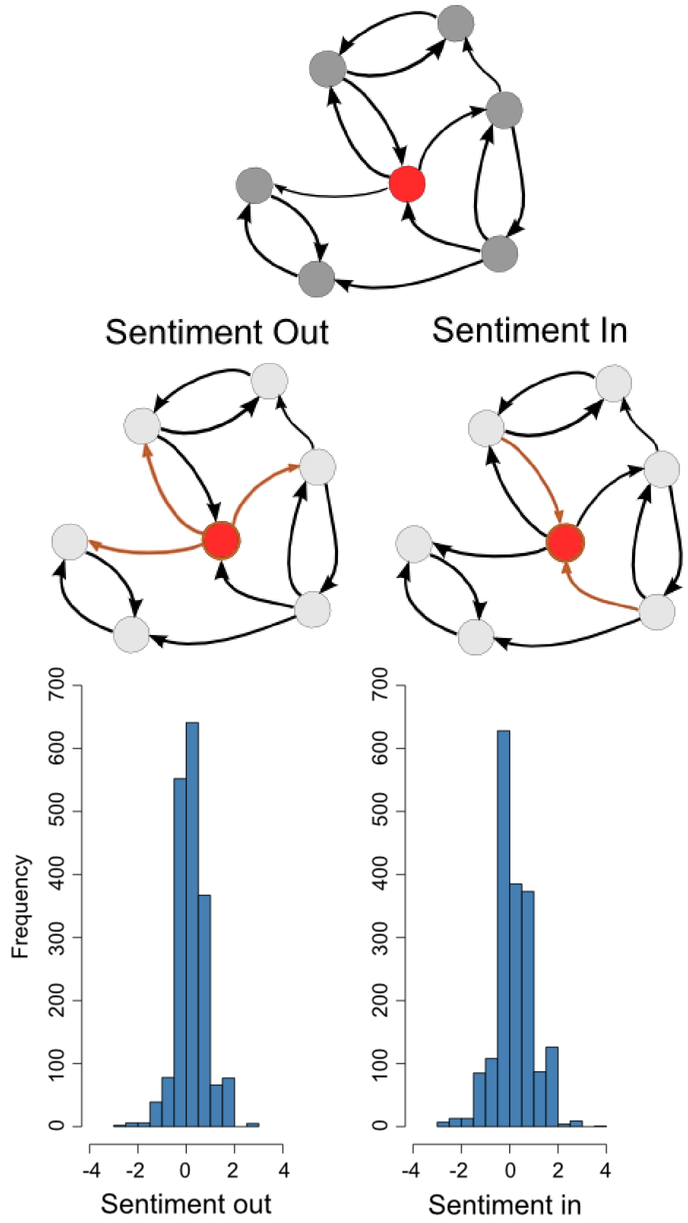
Where we left off with the network

	Mention network		Follower network	
	Full	Reciprocal	Full	Reciprocal
Nodes	40,812	2,047	36,674	2,047
Links	227,203	69,022	3,309,687	173,137
Reciprocal links	23,713	22,218	1,398,236	85,986
Avg. out degree	9	34	90	85
Transitivity	0.02	0.13	0.09	0.28



Tweet sentiment

- Weight mentions network by sentiment
- Calculate some nodal statistics
  - Average sentiment out
  - Average sentiment in
    - Correlation 0.59
- How could these sentiment scores fail?
    - People may tweet out random nonsense
    - SentiStrength may be inaccurate
  - Given sentiment distribution & network topology
    - How likely are we to observe
      - Correlation
      - Connectivity between users

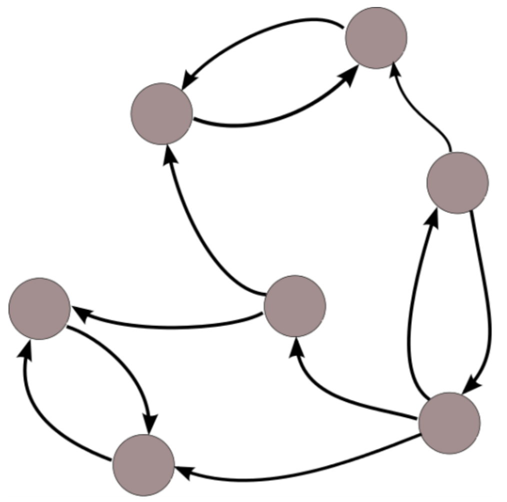


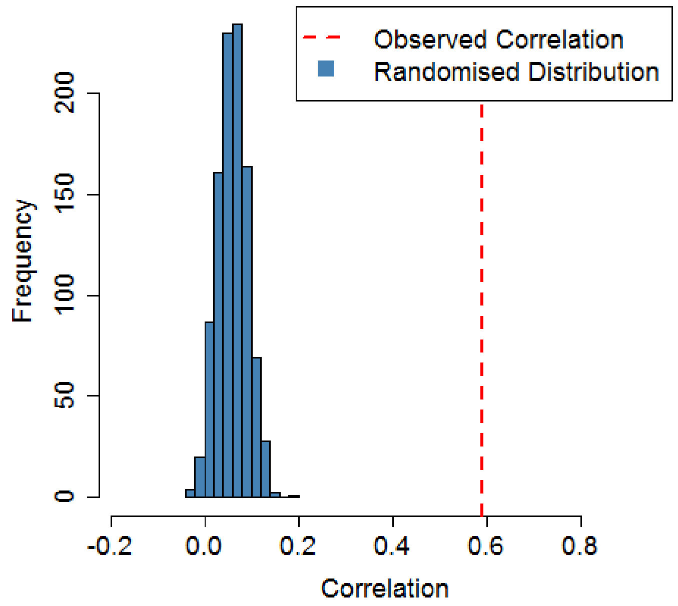
Randomisation tests

- Correlation between sentiment in and out
  - Hold the network topology constant
  - Draw samples (with replacement)
  - Recalculate the correlation
    - Repeat M times
  - Compare randomised distribution to the observed correlation

  - You get what you receive – kinda

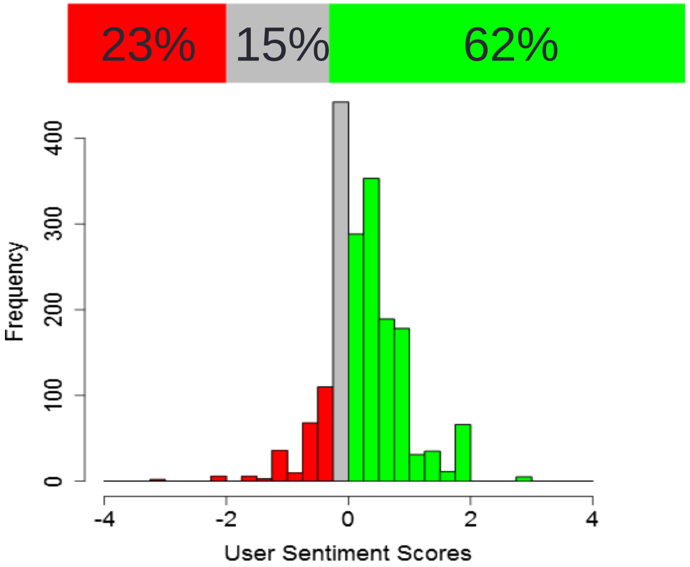
- What about the connectivity patterns?
  - (and not forgetting the follower network!)

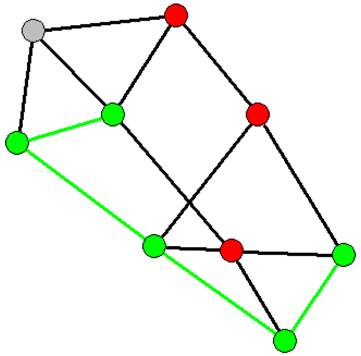




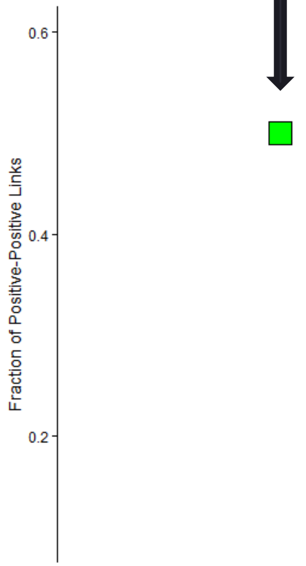
Sentiment & Homophily

- Examine connectivity patterns





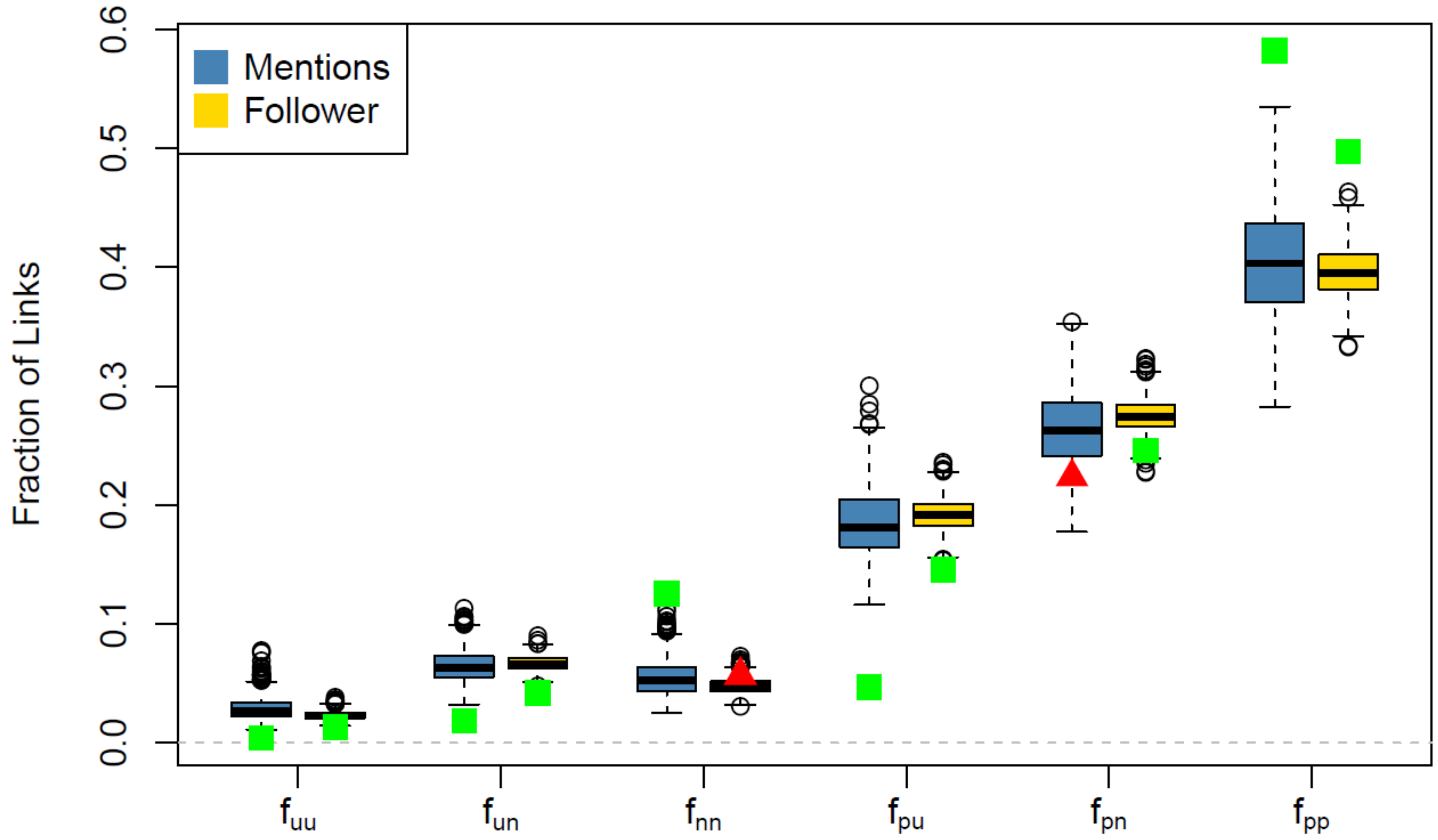
Observed fraction of P-P links



Simulated from randomised network labels

- All connection types
- Both networks

Randomisation tests



- Sentiment: correlation and clustering
  - Proxy for homophily, but still noisy…
- Can we use this with groups of yes and no voters?

Again, another little break with R

- So what did this analysis actually look like…

- 3_homophily_sentiment

COMMUNITY DETECTION