SICSS-Edinburgh

Introduction to Computational Social Science

Björn Ross

14 June 2022

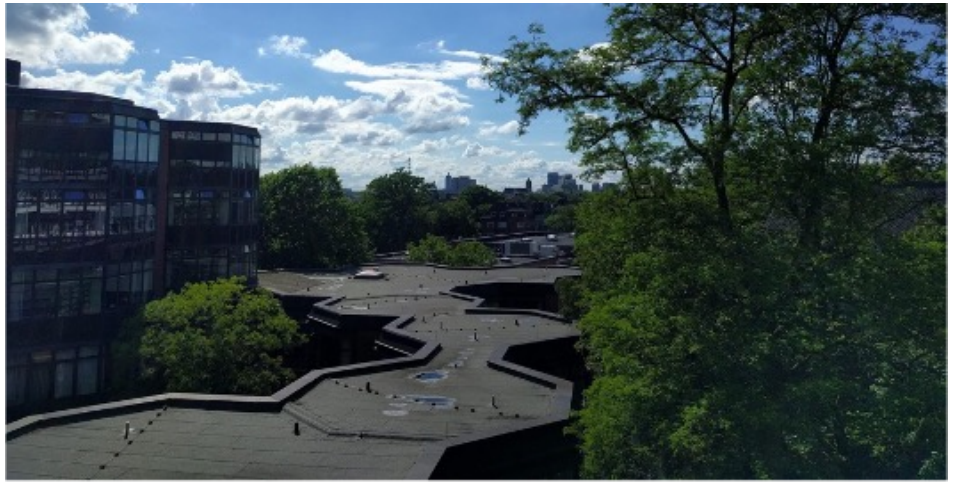
Introduction

Background

University of Münster

- B.Sc. in Information Systems
  - Dissertation on comparing sentiment analysis methods
- M.Sc. in Computer Science
  - Dissertation on predicting the box office revenue of films based on tweets

Background



University of Duisburg-Essen

- PhD on “The Diffusion of Emotions, Information and Opinions on Social Media”
    - Aspects of information systems, computer science, social science

Research Overview

Research Interests

- Negative sides of social media
  - Automated communication (“social bots”)
  - Fake news
  - Hate speech
- Using social media for social good (e.g. crisis communication)
- Methods to study social media (and related challenges)
- Looking beyond social media

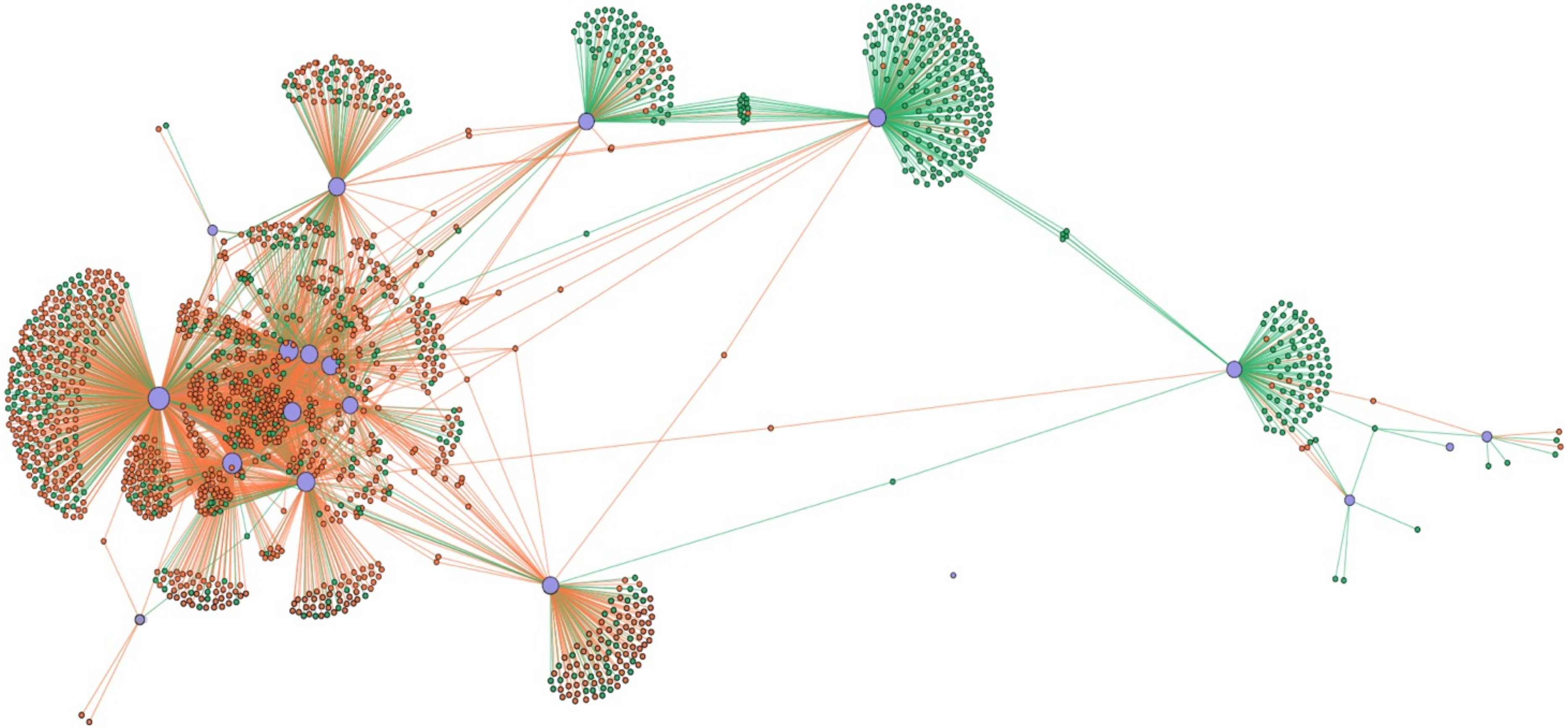
SICSS-2022

Agenda

"Yesterday you learned how to get the data – now what do you do with it?"

- 10:00-13:00 Quantitative and computational approaches (Björn Ross)
  - Social network analysis
    -> Wednesday (Tod Van Gunten)
  - Computational text analysis and natural language processing
    -> Thursday (Chris Barrie and Walid Magdy)
  - Machine learning and prediction
    -> Friday (Walid Magdy and Björn Ross)
  - Statistical analysis of social media data
  - Agent-based simulation models
- 14:00-16:00 Digital qualitative methods (Karen Gregory)

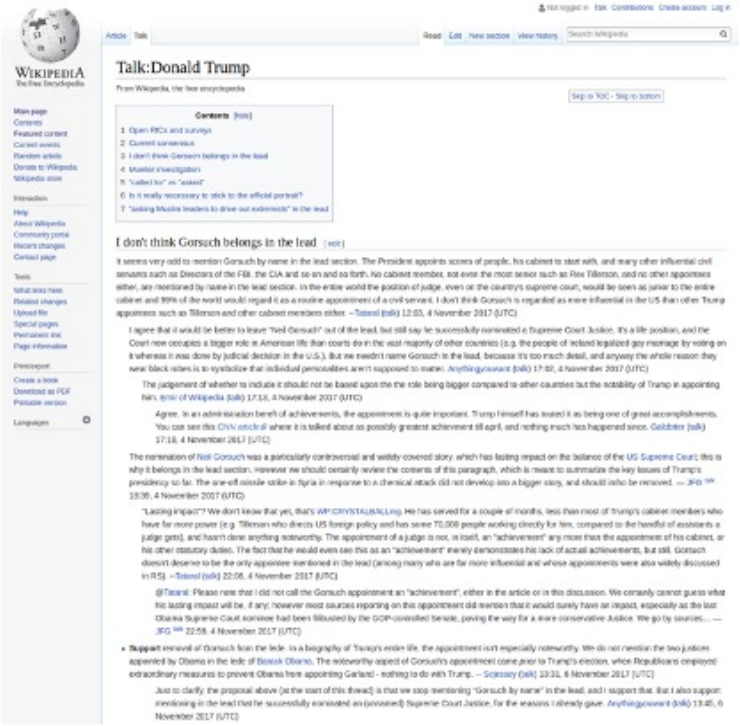
Social Network Analysis



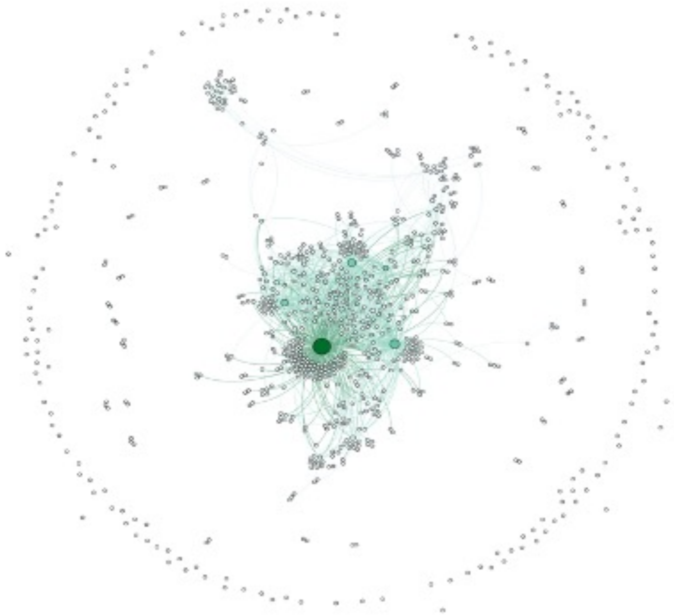
25 most commented tracks on SoundCloud

Ross et al. (2018). Social bots in a commercial context-A case study on SoundCloud. In Proceedings of ECIS 2018.

Parsing Wikipedia data into networks



Grawitas



Conversation Network

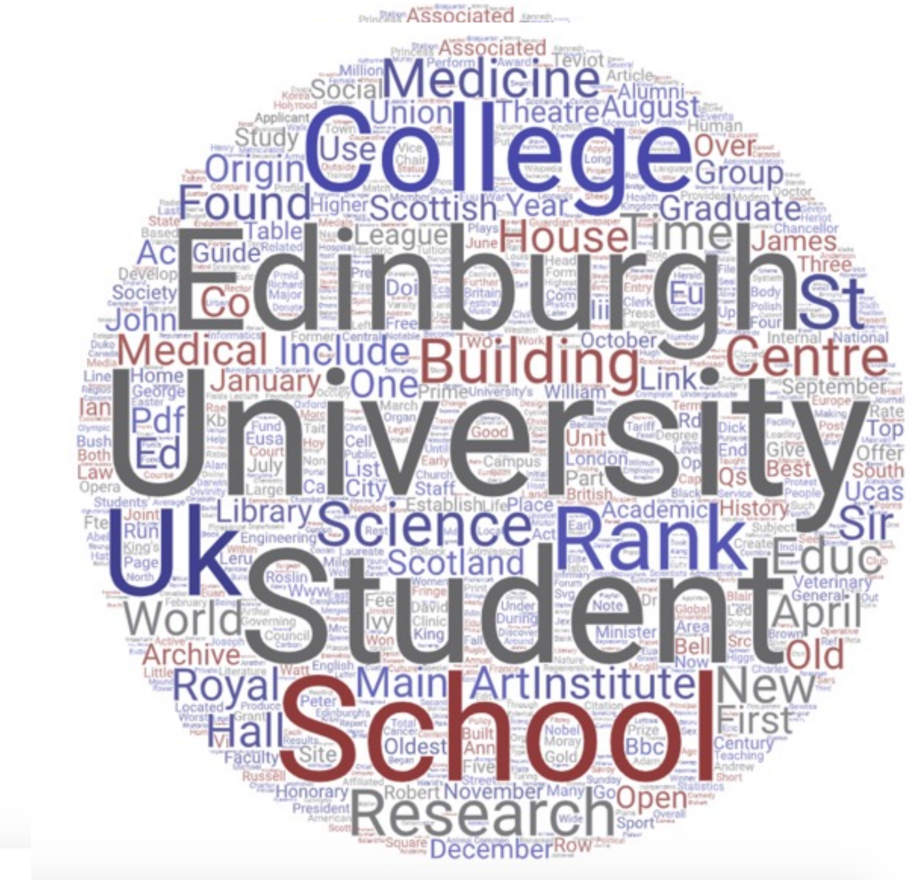
https://github.com/bencabrera/grawitas

Computational Text Analysis and Natural Language Processing

Word frequency analysis

- Very simple starting point

1. Preprocess date (lowercasing?...)
2. Count words
3. Normalize by document length
4. Average across all documents



Dictionaries and lexicons

- What if we know what we are looking for?
- Dictionaries (lexicons) are prebuilt mappings
  - Category -> word list
  - E.g., a tiny sentiment lexicon:
    - Positive: good, great, happy, amazing, wonderful, best, incredible
    - Negative: terrible, horrible, bad, awful, nasty, gross, worst, poor

- Domain can be important
  - "*unpredictable* movie plot"
  - "*unpredictable* coffee pot"

Some example dictionaries

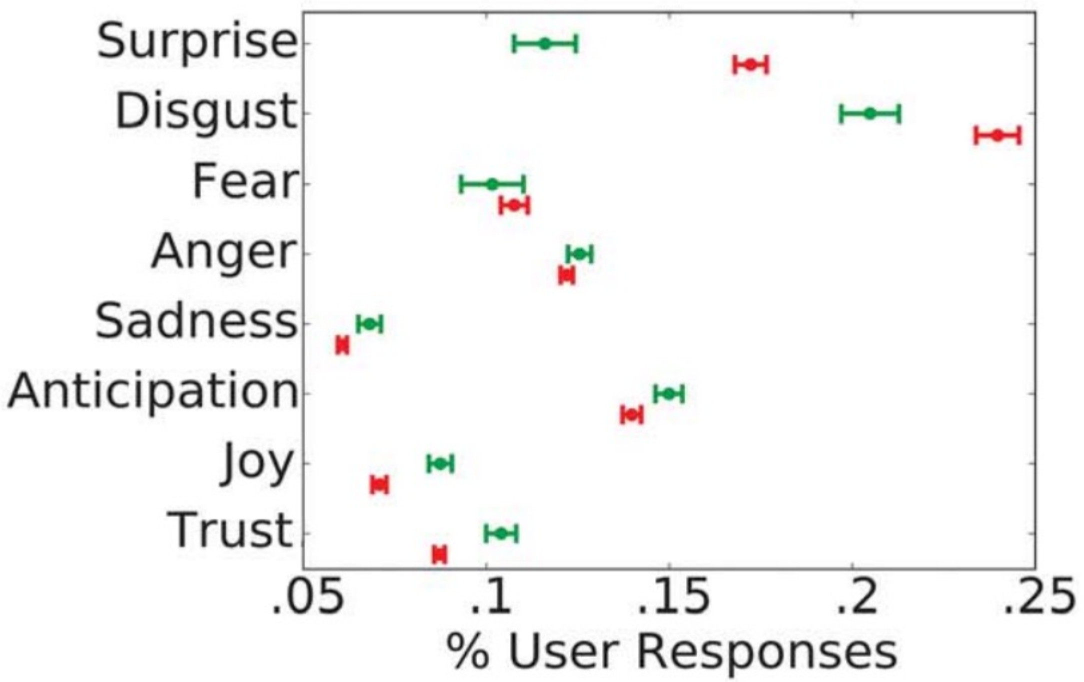
- LIWC (Pennebaker et al. 2015)
- General Inquirer (Stone 1997)
- Roget's Thesaurus Categories
- VADER (Hutto and Gilbert, 2014)
- Sentiwordnet (Esuli and Sebastiani 2006)
- Wordnet Domains (Magnini and Cavaglia, 2000)
- EmoLex (Mohammad and Turney, 2010)
- Empath (Fast et al., 2016)
- Personal Values Lexicon (Wilson et al., 2018)
- …

How to get a score per category?

$$\frac{num\_dictionary\_words\_in\_document}{num\_total\_words\_in\_document}$$

- That's it!
- Can also be used as input in machine learning

Reactions to rumour tweets with EmoLex



Red = reactions to false rumors
Green = reactions to true rumors

Vosoughi, Roy, and Aral, 2018

LIWC category dominance scores

Truthful			
Interviews		Trials	
Class	Score	Class	Score
Metaphor	2.98	You	3.99
Money	2.74	Family	3.07
Inhibition	2.74	Home	2.45
Home	2.13	Humans	1.87
Humans	2.02	Posemo	1.81
Family	1.96	Insight	1.64

Deceptive			
Interviews		Trials	
Class	Score	Class	Score
Assent	4.81	Anger	2.61
Past	2.59	Anxiety	2.61
Sexual	2.00	Certain	2.28
Other	1.87	Death	1.96
Motion	1.68	Physical	1.77
Negemo	1.44	Negemo	1.52

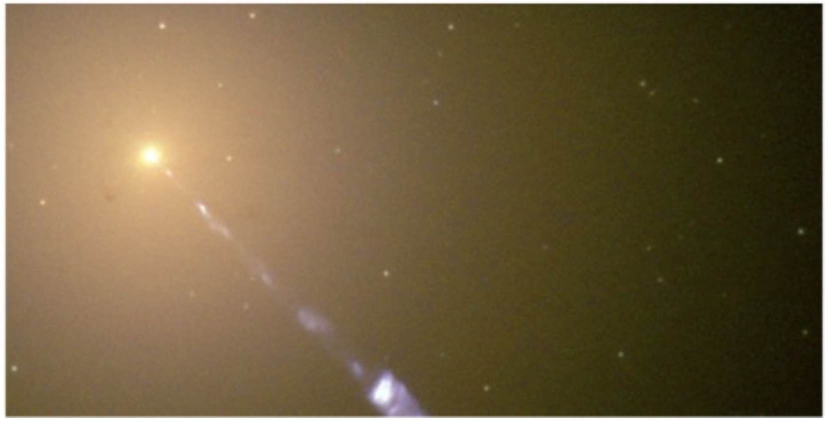
Pérez-Rosas et al, 2015

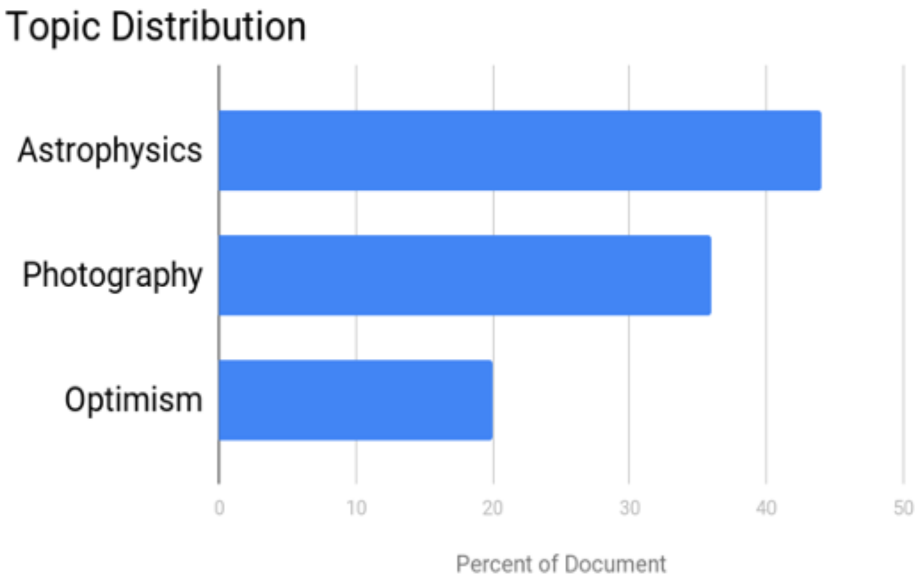
Topic modelling

The New York Times

Expected Soon: First-Ever Photo of a Black Hole

Have astronomers finally recorded an image of a black hole? The world will know on Wednesday.

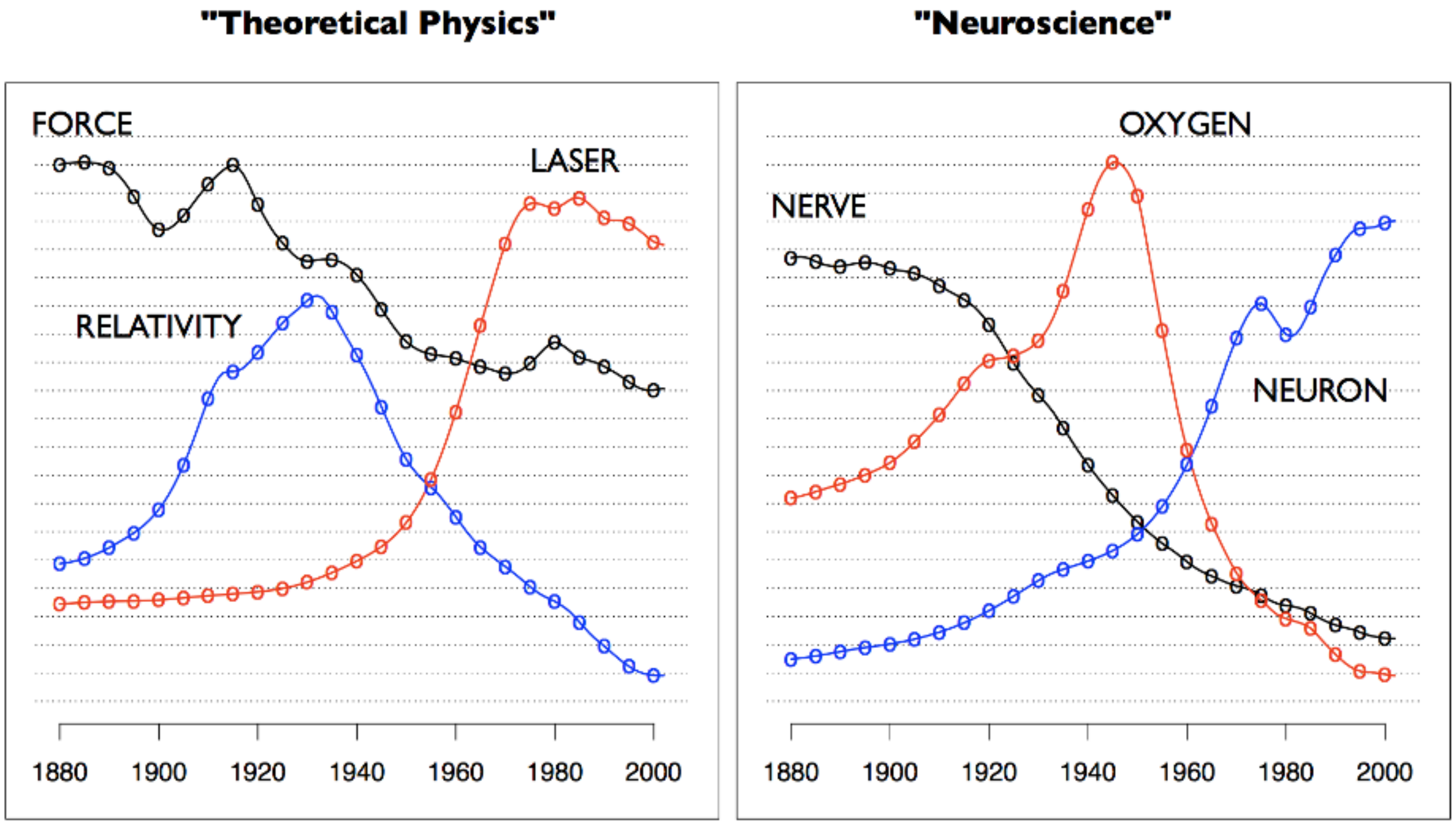




Topic modelling

human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Topic modelling



Machine learning approaches

Sentiment analysis:

Features

Target variable

	Contains word *dog*	Contains word *cat*	Contains word *cute*	Contains word *ugly*	Sentiment
“That dog is cute”	yes	no	yes	no	positive
“That cat is cute”	no	yes	yes	no	positive
“That dog is ugly”	yes	no	no	yes	negative
“That cat is ugly”	no	yes	no	yes	negative
“That unicorn is cute”	no	no	yes	no	?

Combining methods

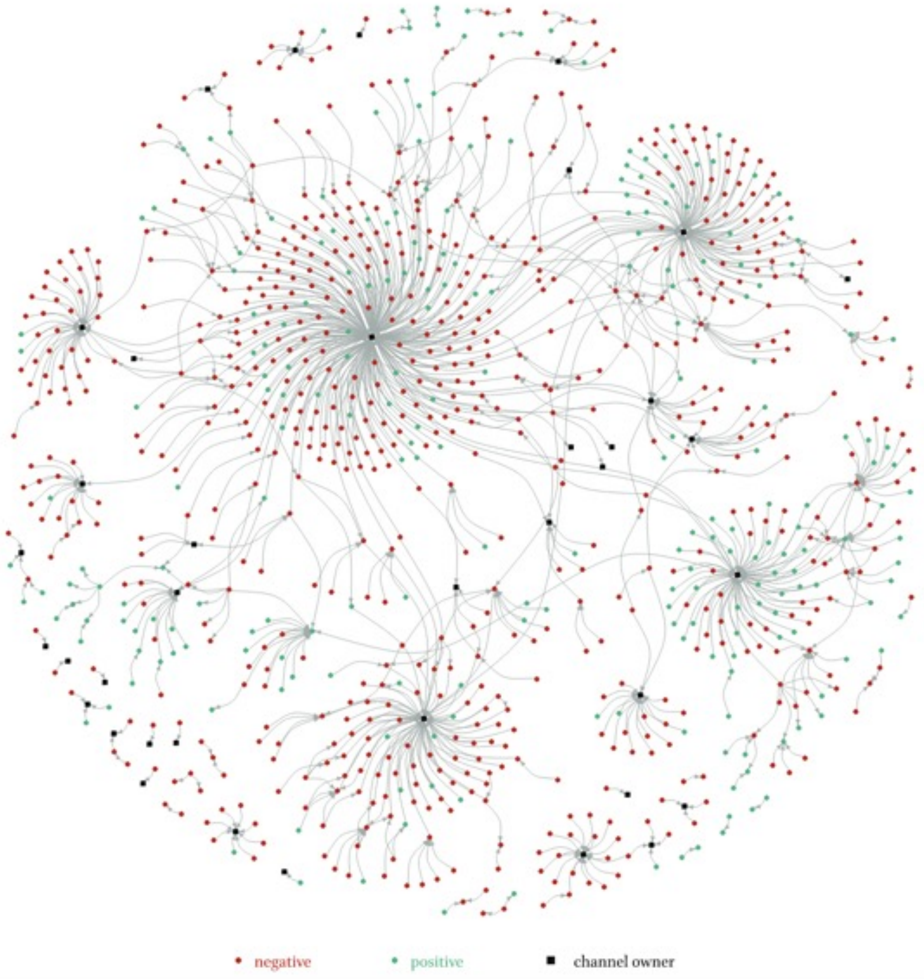
Search keyword	total results	total likes	total dislikes	total views	total comments	filtered comments
Adoption for same-sex couples	266	31,876	8,509	2,576,318	15,889	8,443
Headscarf ban in Germany	320	199,912	26,393	7,247,958	48,354	14,277
Climate change	336	167,236	16,136	10,387,029	46,894	18,185

- Crawled YouTube videos
- Annotated random sample of 4,000 comments (two annotators, Krippendorff‘s α between 0.54 and 0.67)

	Dataset		
Sentiment	Adoption rights	Headscarf ban	Climate change
Negative	339	400	416
Positive	530	294	356
Others	2432	2769	2328

- Trained machine learning-based classifier to detect opinions on entire dataset
- Visualised network and calculated network statistics

Dataset	Sub-network	Sentiment	Statistics			
			Internal Ties	External Ties	Class E-I Index	Global E-I Index
Adoption rights	I	Negative	1	15	0.88	0.92
		Positive	2	58	0.93	
	II	Negative	9	19	0.36	0.61
		Positive	0	18	1	
	III	Negative	1	12	0.85	0.94
		Positive	0	22	1	
Headscarf ban	I	Negative	30	182	0.72	0.77
		Positive	0	49	1	
	II	Negative	28	71	0.43	0.59
		Positive	1	40	0.95	
	III	Negative	42	71	0.26	0.35
		Positive	0	17	1	
Climate change	I	Negative	2	48	0.92	0.89
		Positive	3	39	0.86	
	II	Negative	8	35	0.63	0.79
		Positive	1	41	0.95	
	III	Negative	4	26	0.73	0.61
		Positive	11	36	0.53	



Headscarf ban discussion network

Toolbox metaphor

Computational text analysis and natural language processing

Social network analysis

Machine learning

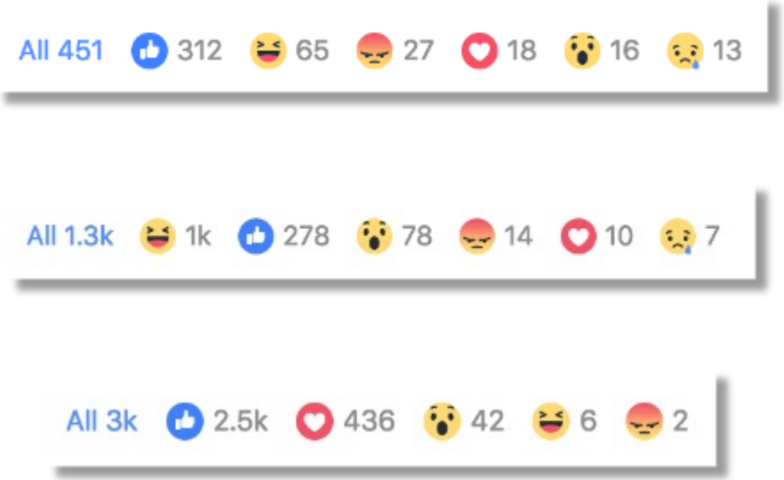
Agent-based simulation modelling

Statistical analysiss

“if all you have is a hammer, everything looks like a nail”

Statistical analysis of social media data

Let’s start with an example…



(Ross et al. 2018)

Context

- Most research focuses on Twitter (Munro & Manning 2012) although it is considered an 'elite channel' in times of crisis (Eriksson & Olsson 2016)
- Facebook makes it possible to study reactions in more detail (👍 / 😢 / 😠 / 😆 / 😮 / ❤️)

- Research questions
  - How should crisis-related information be published on Facebook to reach as many people as possible?
  - What is the relationship between the content of a post and the type of reaction it will receive?

Research Design

- Data collection
  - Posts and reactions from six Facebook pages

	Berlin	London	Stockholm
Time span	19-26 Dec 2016	22-29 Mar 2017	7-14 Apr 2017
Municipality	Berlin.de	London Gov	Stockholms stad
Emergency service agency	Polizei Berlin	Metropolitan Police Service	Krisinformation.se

Research Design

- Data preparation
  - Annotation of posts: categories from Ehnis et al. (2014)
    - Information
      - Number of victims
    - Encouragement of behaviour
    - Warnings
    - Condolences
  - Cleaning of data set, e.g.
    - Posts not about the crisis excluded
    - Pages with very few posts (London Gov, Berlin.de) excluded
    - Information excluded as a category due to low reliability
  - Final data set: 66 posts by four Facebook pages

Research Design

- Data analysis
    - Negative binomial regression (log link)
        - Dependent variables: Number of shares, number of reactions
        - Independent variables:
            - Text length
            - Presence/absence of image
            - Presence/absence of video
            - Category of post content
        - Control variable: Page that the post appeared on
    - Correlation coefficients between reaction types

Results: Number of interactions

- Regression results:

	Number of shares					Number of reactions				
Variable	β	exp(β)	SE	Z	p	β	exp(β)	SE	Z	p
(Intercept)	5.20		1.30	4.02	$< .000^*$	3.72		0.86	4.32	$< .000^*$
Controls										
Page Krisinformation.se	2.57	13.10	0.58	4.43	$< .000^*$	2.91	18.35	0.39	7.51	$< .000^*$
Page Metropolitan Police London	3.44	31.09	0.65	5.32	$< .000^*$	3.50	33.21	0.43	8.18	$< .000^*$
Page Polizei Berlin	5.53	251.03	0.73	7.59	$< .000^*$	5.06	158.36	0.48	10.47	$< .000^*$
Explanatory variables										
log(Text length)	−0.39	0.68	0.15	−2.55	$.011^*$	−0.14	0.87	0.10	−1.42	.157
Image	0.47	1.59	0.56	0.83	.405	0.77	2.16	0.37	2.07	$.038^*$
Video	1.22	3.40	0.88	1.40	.163	1.36	3.89	0.58	2.33	$.020^*$
Number of victims	0.59	1.81	0.76	0.78	.434	0.59	1.81	0.50	1.18	.238
Warning	1.00	2.71	0.76	1.32	.188	0.43	1.53	0.50	0.85	.398
Encouragement	0.37	1.44	0.44	0.84	.404	−0.21	0.81	0.29	−0.71	.478
Condolences	0.50	1.65	0.52	0.97	.333	1.10	2.99	0.34	3.19	$.001^*$

$^*p < .05$

- Shorter posts are shared more often
- Posts with image, video, condolences receive more reactions

Results: Reaction types

- Correlation matrix of reactions:

	Likes	Sadness	Angry	Wow	Haha	Love
Shares	.785	.731	.334	.469	.275	.526
Likes		.935	.149	.426	.175	.859
Sadness			.342	.473	.041	.748
Angry				.639	.224	.009
Wow					.526	.210
Haha						.002

- Negative emotions predominate, esp. early on
- Positive emotions still present, especially later
  - '... we have been overwhelmed by the love and support for our family, and most especially, the outpouring of love and respect for our Keith . . . ' (26 Mar 2017, 1:37 PM)

Results: Reaction types

- Page owners can gauge the reception of a post based on Facebook ‘Reactions’ feature
- Recommendations to page owners can be made that are backed up by empirical evidence, e.g.
  - Keep your posts concise – text length influences shares
  - Use image and/or video – it boosts reactions
- Posts with condolences receive many reactions → problematic paradox
  - ESAs may want informational posts to diffuse through the network
  - Audiences seem to favour emotional posts

Statistical analysis of social media data

What’s different in computational social science?

Things to consider

- (Sometimes) different statistical models
  - Count data
  - Time series analysis
- (Sometimes) different unit of analysis compared with e.g. psychology
  - Posts
  - Organisations
- (Sometimes) different order of magnitude of data
  - Thousands of accounts?
  - Millions of posts?
- Other differences?

Things to consider

Experimental evidence of massive-scale emotional contagion through social networks

Adam D. I. Kramer[a,1], Jamie E. Guillory[b,2], and Jeffrey T. Hancock[b,c]

[a]Core Data Science Team, Facebook, Inc., Menlo Park, CA 94025; and Departments of [b]Communication and [c]Information Science, Cornell University, Ithaca, NY 14853

emotional post omitted from their News Feed on a given viewing, such that people who had more content omitted were given higher weight in the regression. When positive posts were reduced in the News Feed, the percentage of positive words in people's status updates decreased by $B = -0.1\%$ compared with control [$t(310{,}044) = -5.63$, $P < 0.001$, Cohen's $d = 0.02$], whereas the percentage of words that were negative increased by $B = 0.04\%$ ($t = 2.71$, $P = 0.007$, $d = 0.001$). Conversely, when negative posts were reduced, the percent of words that were negative decreased by $B = -0.07\%$ [$t(310{,}541) = -5.51$, $P < 0.001$, $d = 0.02$] and the percentage of words that were positive, conversely, increased by

TECH

Facebook Manipulated 689,003 Users' Emotions For Science

Kashmir Hill Former Staff
Welcome to The Not-So Private Parts where technology & privacy collide

Follow

Jun 28, 2014, 02:00pm EDT

This article is more than 7 years old.

June 29: Updated with statement from Facebook, from the author of the study, and from the editor of the academic journal that published the study.

Facebook is the best human research lab ever. There's no need to get experiment participants to sign pesky consent forms as they've already agreed to the site's data use policy. A team of Facebook data scientists are constantly coming up with new ways to study human behavior through the social network. When the team

st

The Atlantic

TECHNOLOGY

Everything We Know About Facebook's Secret Mood-Manipulation Experiment

It was probably legal. But was it ethical?

By Robinson Meyer

MICHELLE N. MEYER OPINION JUN 30, 2014 3:22 PM

Everything You Need to Know About Facebook's Controversial Emotion Experiment

Facebook conducted a study for one week in 2012 testing the effects of manipulating News Feed based on emotions. The results have hit the media like a bomb. What did the study find? Was it ethical? And what could or should have been changed?

Things to consider

- Ethical issues with manipulating variables of interest
- (Sometimes) very high sample sizes – “Too Big to Fail” models
  Related: statistical significance vs. practical significance (Lin et al. 2013)
  - Present effect sizes!
  - Report confidence intervals!
  - Use charts!

Home > Information Systems Research > Vol. 24, No. 4 >

Research Commentary—Too Big to Fail: Large Samples and the *p*-Value Problem

Mingfeng Lin, Henry C. Lucas Jr, Galit Shmueli

Published Online: 22 Oct 2013 | https://doi.org/10.1287/isre.2013.0480

Abstract

The Internet has provided IS researchers with the opportunity to conduct studies with extremely large samples, frequently well over 10,000 observations. There are many advantages to large samples, but researchers using statistical inference must be aware of the *p*-value problem associated with them.

Things to consider

- Population bias
- Proxy population mismatch
- Proprietary algorithms
- Digital divide between academic and industry research
- Relationship between human behaviour and online platform design
  - Distortion of human behaviour by platform-specific features
- Incomparability of methods and data, lack of good benchmarking data
- Multiple comparisons problems, multiple hypothesis testing

Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. Science, 346(6213), 1063-1064.

Group activity

- Form small groups
- Write down a research question that you are interested in
  (are researching / are planning to research)
- Look at other group member's questions
  - Which, if any, methods would you use to address them?
  - Why or why not?
    Which methods did you consider and which issues with them did you see?
- Ask the person who wrote down
- Which methods are you best at? Which ones do you usually apply?
- Later: One member reports to plenary
  1-2 minutes per group

Reports from group activity

Agent-based simulation models

Social bots

Intent (Ferrara et al. 2016)				
Imitation of human behaviour (Boshmaf et al. 2013)		**Malicious**	**Neutral**	**Benign**
	High: Social bots	Astroturfing bots (Ratkiewicz et al., 2011)	Humoristic bots (Veale et al., 2015)	Chat bots (Salto Martínez & Jacques García, 2012)
	Low to none	Spam bots (Wang, 2010)	Nonsense bots (Wilkie et al., 2015)	News bots (Lokot & Diakopoulos, 2016)

Are social bots a real threat?

Background

- Spiral of silence theory (Noelle-Neumann, 1974)
  - Explains how public opinion forms
  - Individuals
    - fear isolation
    - keep track of the opinions of others on contentious issues
    - Become less (more) likely to express their opinion if they perceive themselves to be in the minority (majority)
  - Groups
    - Converge towards a consensus over time as (perceived) minority is silenced
    - May ultimately accept one opinion as the (perceived) majority opinion

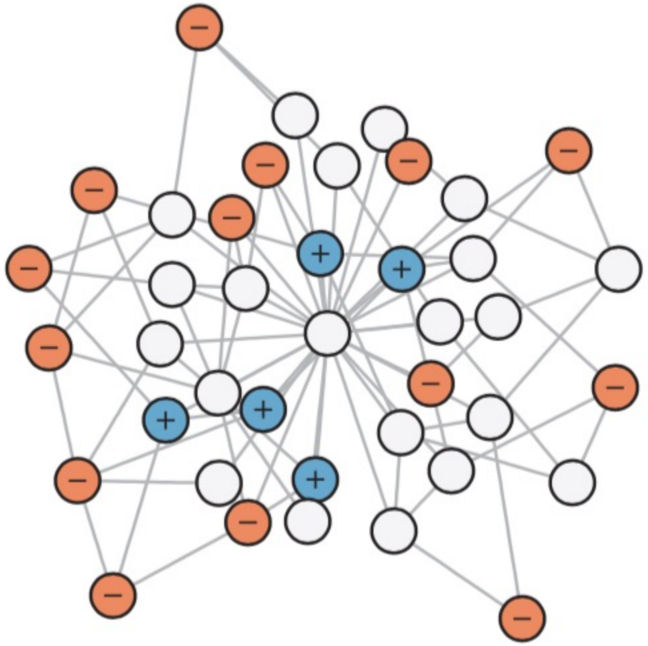
Method

- Simulation software NetLogo (Wilensky, 1999)
- Agents are connected in a network (Barabási & Albert, 1999; Dorogovtsev et al., 2000)
- Agent *i*'s variables: (see also Sohn & Geidner, 2016)
  - Fixed opinion o_i in {+,–}
  - Fixed willingness to self-censor Φ_i in [0; 1] (Hayes et al., 2005)
  - Confidence $c_i(t)$ in [0; 1]
  - $c_i > \Phi_i$ means that agent speaks out
  - Confidence changes over time as agent observes its neighbours in the network (= diffusion of opinions)

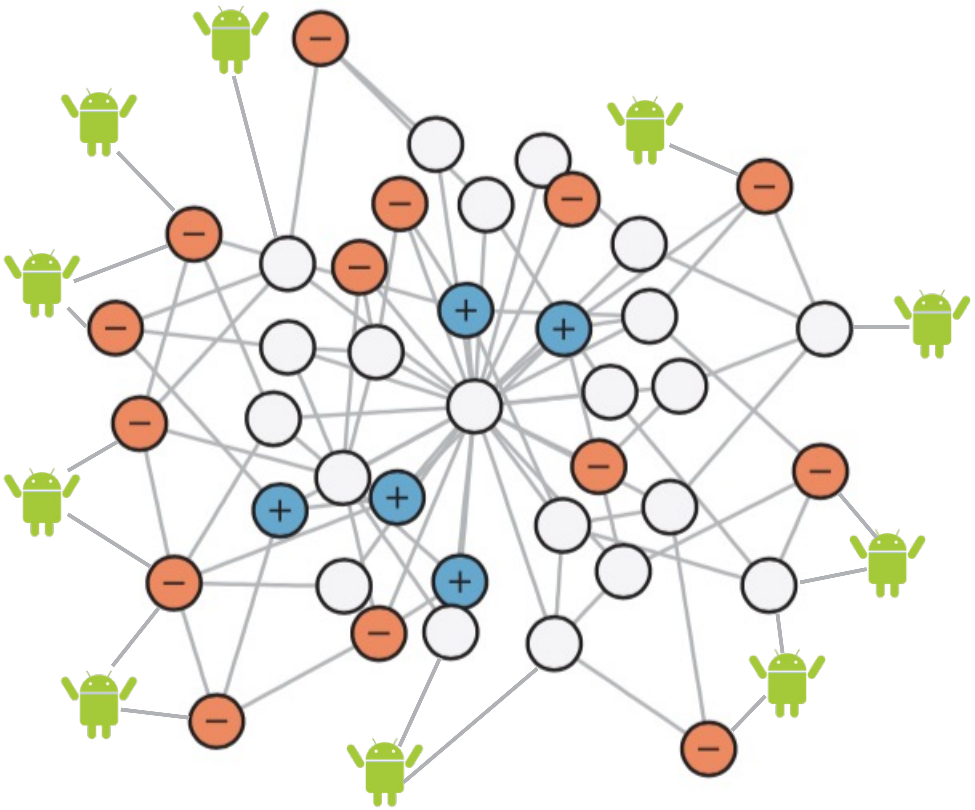
Method: Confidence changes

- How does confidence $c_i(t)$ change over time depending on agent's environment?

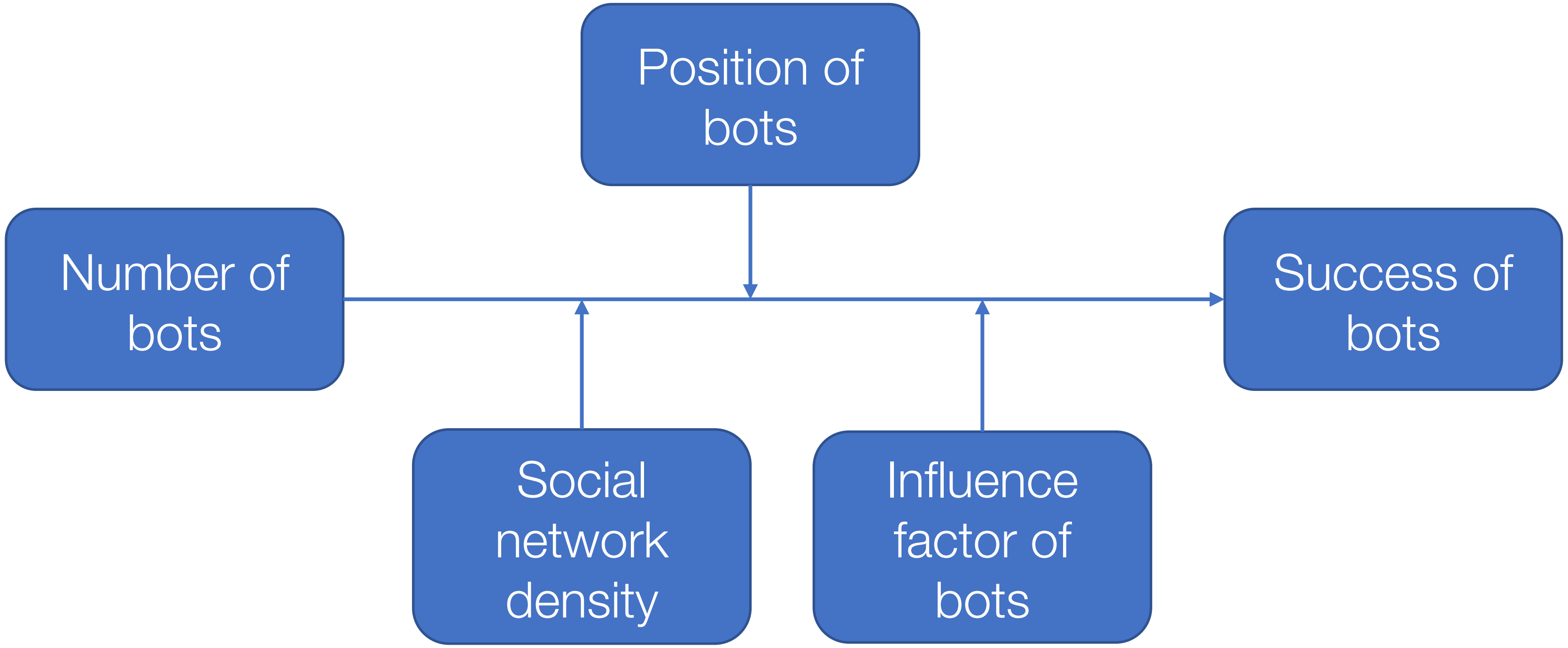
Agent-based model



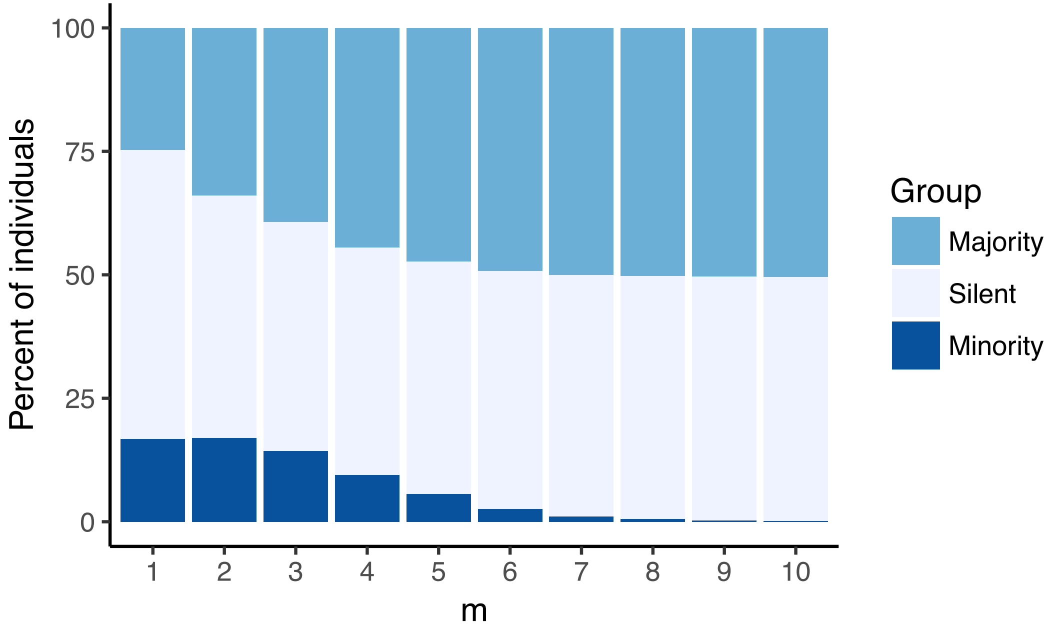
Agent-based model



Research model

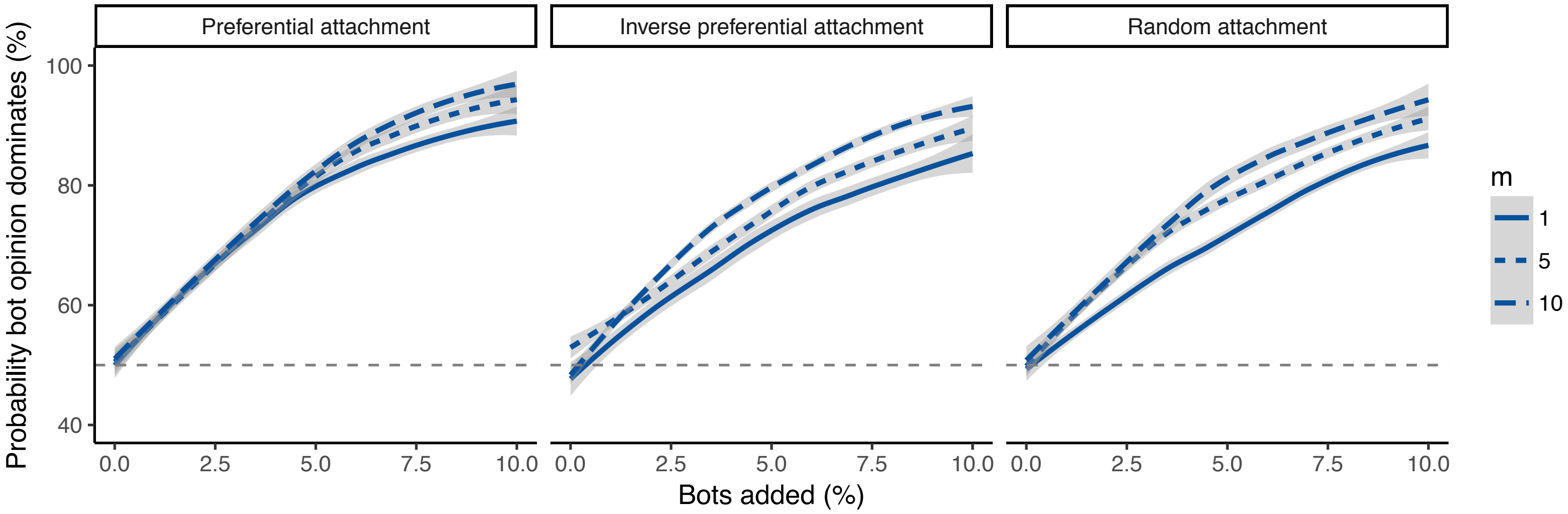


Results



Effects of network density on dominance of majority over minority opinion

Results



Influence of bots added at different positions in the network

Implications

- Model allows us to observe the “bridge” from individual to group behaviour
  - Network model differs from previous research
- Plausible mechanism of manipulation on the basis of an established theory of opinion formation
  - Potential threat to decision-making processes
- Simplifying assumptions (e.g. bots only supported one opinion)
- Open questions: e.g. regulation

When is ABM useful?

Group discussion

Questions and Discussion

University of Edinburgh

Björn Ross

14 June 2022