SAND2022-12497C

Sandia National Laboratories

Exceptional service in the national interest

A Complex, Integrative Agent-Based Model of Disinformation Cascades.

Matthew Sweitzer, 1462

September 20th, 2022

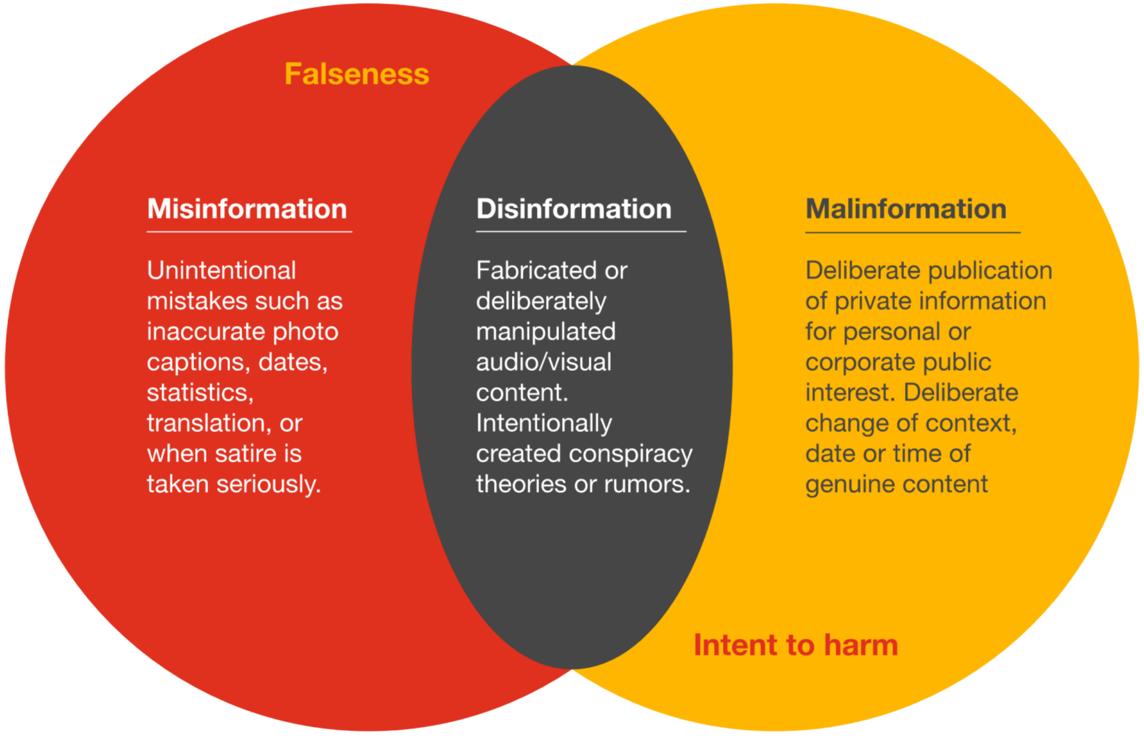
SBB-BRIMS 2022

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

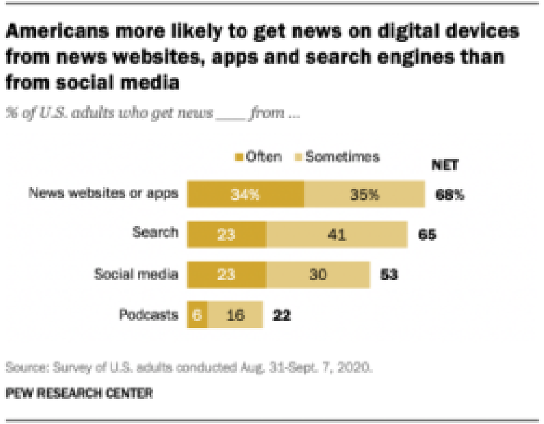
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Disinformation is being used by many nation-states

- Disinformation is false information intentionally used for harm.
  - Nation-State and non-state actors use disinformation.
  - Social media platforms a means of disseminating disinformation.
- Machine Learning/Artificial Intelligence techniques for:
  - Identifying false information.
  - Predicting the spread of information.
  - Predicting who will adopt information.
- However:
  - Complex social system with many interacting factors.
  - Adversaries are changing tactics.
  - We can’t (ethically) experiment with the real world.
  - We have limited ground truth.
  - Environment is changing.
    - Dataset shift problem.

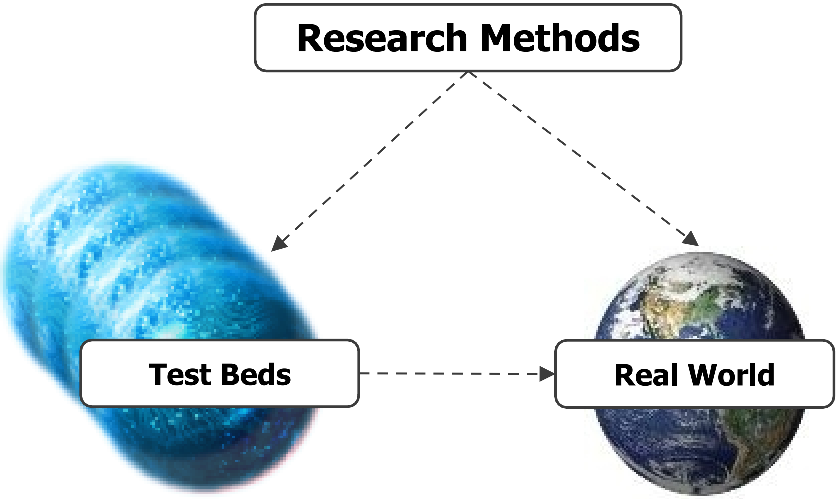


Source: FirstDraft, The essential guide to understanding the information disorder, 2019.



We are investigating the use of social simulations as a testbed.

- Our approach: Use social simulations as a proxy for the real world.
- Social simulations are computational models of real-world phenomena.
  - Methods include agent-based modeling, systems dynamics, ...
- Often used for better understanding a phenomena and testing interventions in a virtual world.
- Simulations can help solve some of the problems:
  - Full ground truth.
  - Can control data bias.
  - Can run experiments and counterfactuals.
  - Can evaluate performance on varied models, parameterizations, etc.

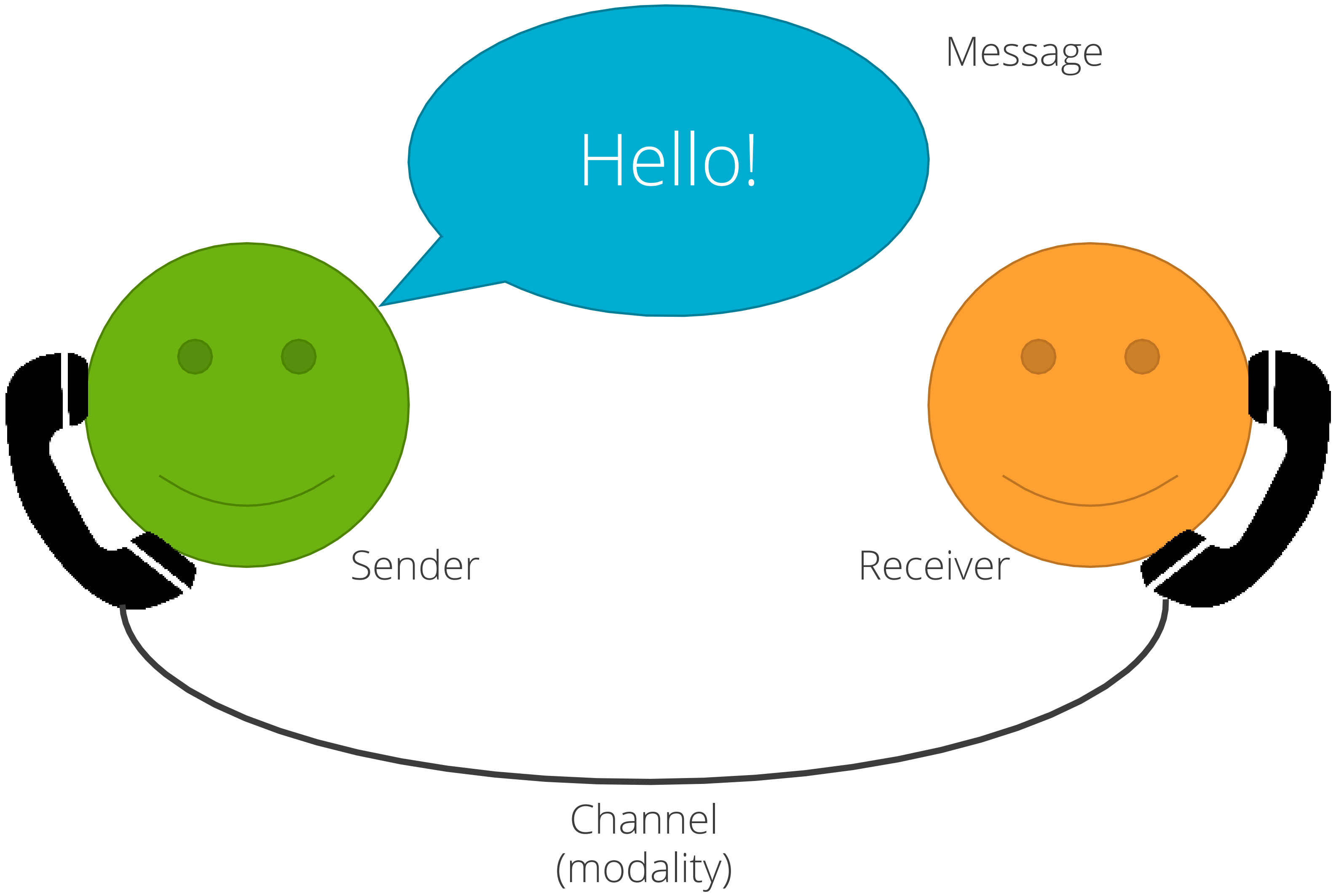


How does the complexity of the environment impact the learnability and generalizability of ML models?

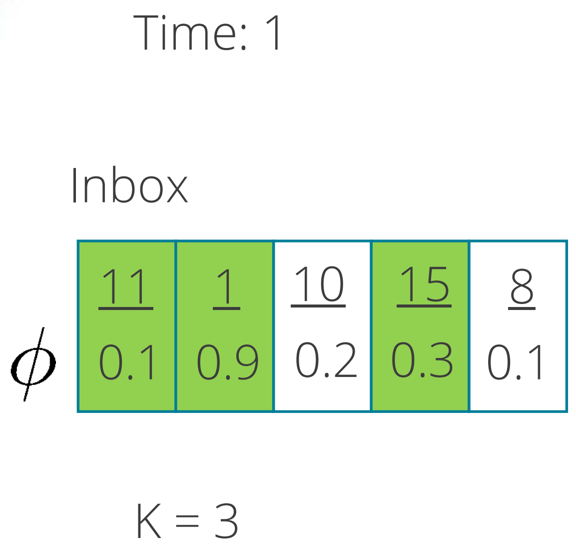
Modeling Process

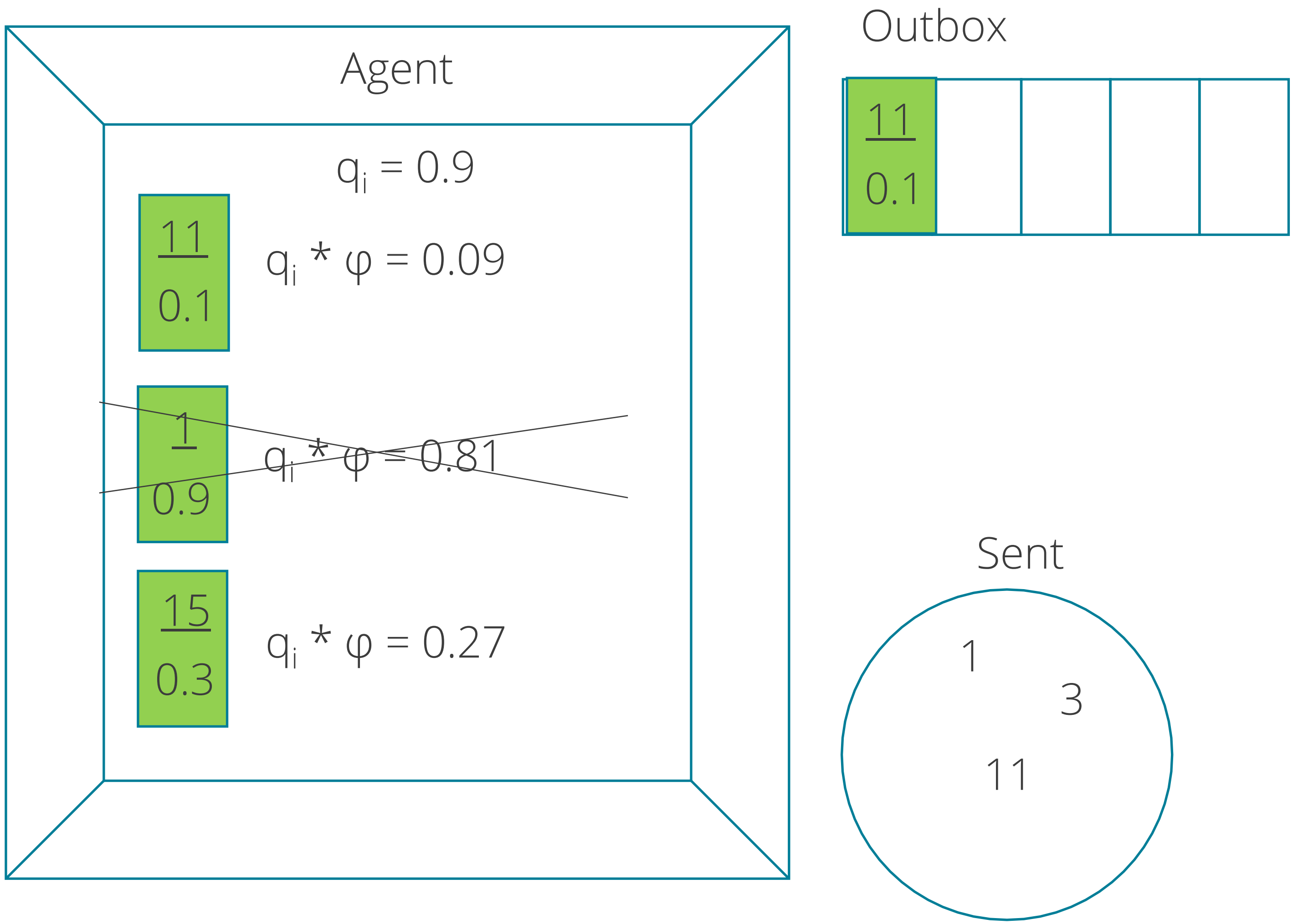
- Create a simple agent-based modeling framework for person-to-person communication to **generate cascade data.**
  - Can adapt to various theoretical additions at the agent-, network-, or message-level
- Challenges:
  - Many different theories from different disciplines apply (social-psychology, communications, group theory, etc.).
  - Most existing simulations (from information diffusion, epidemic modeling) do not generate significant data.
  - Operationalization of multiple theories within the same model.

Berlo (1960) SMCR model of communication



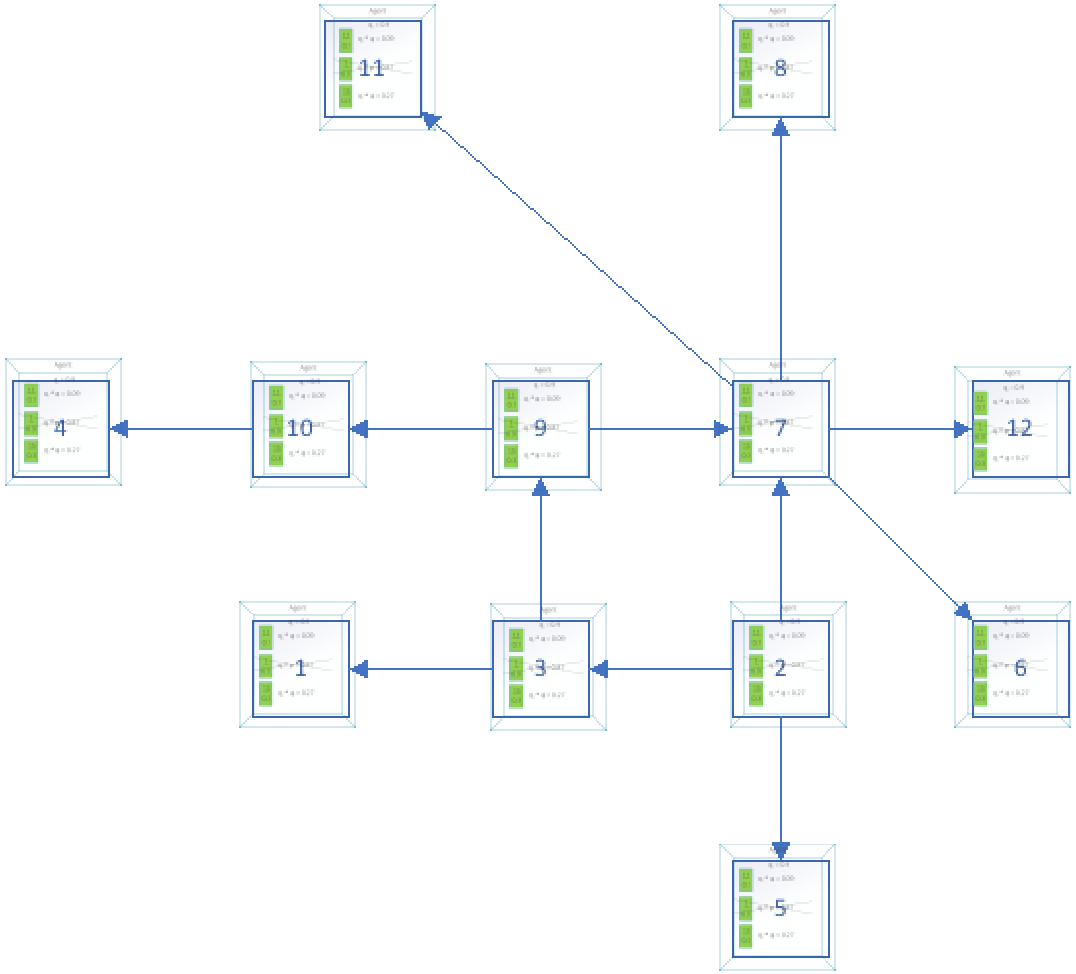
Simple Information Diffusion Model





- Capture attentional constraints (k_i).
- Capture innate virality of messages (φ).
- Captures subjective likelihood to resend (q_i).

Agent model is used for each agent in a social network.



Complex Information Diffusion Model

- **Sender characteristics**
  - Credibility or authority, “speech ability” or persuasiveness, social network centrality, conformity to social norms (i.e., “Spiral of Silence”).
- **Message characteristics**
  - Topic salience, message virality, information accuracy.
- **Channel characteristics**
  - Access to communication modality.
- **Receiver characteristics**
  - Trust, cognitive/ideological consistency, “stubbornness”

Complex Information Diffusion Model – Social Network Centrality

- **Sender characteristic** – a person's "importance" in the network, measured by their connectedness to others
  - A person's centrality is positively related with their influence on others (Ibarra et al., 1993; Kameda et al., 1997; Wang et al., 2015)
  - Centrality is operationalized in ABMs in a wide variety of ways from seeding message (Barbuto et al., 2019), to distinguishing "influencer" agents from a general public (Lotito et al., 2021)
- **In CIDM, centrality acts as a weight on inbox priority** – i.e., compared to other messages received, how likely am I to pay attention to *your* message; or how much does the algorithm weight your message compared to others
  - Eigenvector centrality, rescaled to {0:1}; model-added messages are assigned a value of 2 to ensure they are seen

Complex Information Diffusion Model – Trust

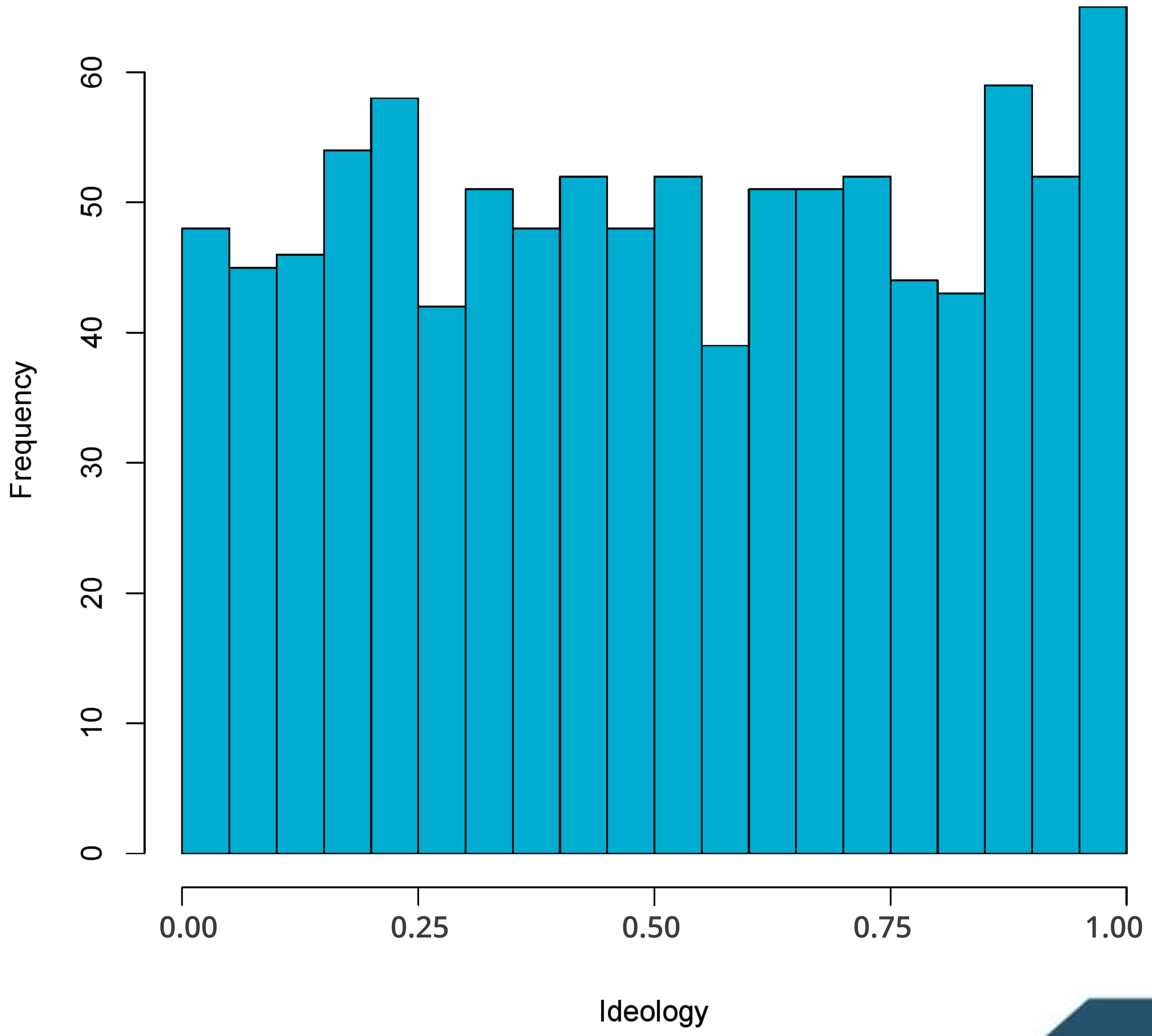
- **Directed receiver-to-sender characteristic** – a person's belief in another that the information they share is true
  - One of many aspects that affects the receiver's perception of the believability of a message, and thereby its adoption and resend probability
  - Commonly implemented as a directed edge weight in the agent-to-agent network affecting adoption and spreading rates (e.g., Hui et al., 2010); less commonly operationalized using tie reciprocity (e.g., Fan et al., 2018)
- **In CIDM, trust is an assigned directed edge value at the start of the model**; not permitted to update in this iteration
  - Can be distributed randomly, as a function of dyadic ideological similarity (Sherchan et al., 2013), or as a function of the proportion of local network overlap (i.e., triadic closure; Igarashi et al., 2008)

Complex Information Diffusion Model – Ideological Consistency

- **Receiver characteristic** – the degree to which the opinion expressed in a message on one topic aligns with the receiver's multi-dimensional ideology; greater similarity increases the probability of adopting the message, and thereby resending
  - Like cognitive dissonance theory (Festinger, 1962), but includes congruency with beliefs on other, related topics
  - Used more often in opinion dynamics models than information diffusion per se (e.g., Lakkaraju, 2016; Schweighofer, 2020)
- **In CIDM, ideological consistency increases resend probability**

Complex Information Diffusion Model – Ideological Consistency

- **Method**
  - Ideology is randomly distributed {0:1}
  - Opinions on some parameterized number of topics are drawn from a gaussian distribution with mean set at ideology, parameterized sd, and opinions beyond 0 and 1 are rounded to floor/ceiling
  - Message asserts some value in opinion space (random; {0:1}) on a particular topic
  - Consistency is 1 – mean distance of message opinion from all non-topic node opinions



Complex Information Diffusion Model

For a message (m), sent by one agent (i) to another (j), the receiving agent will resend the message with the probability:

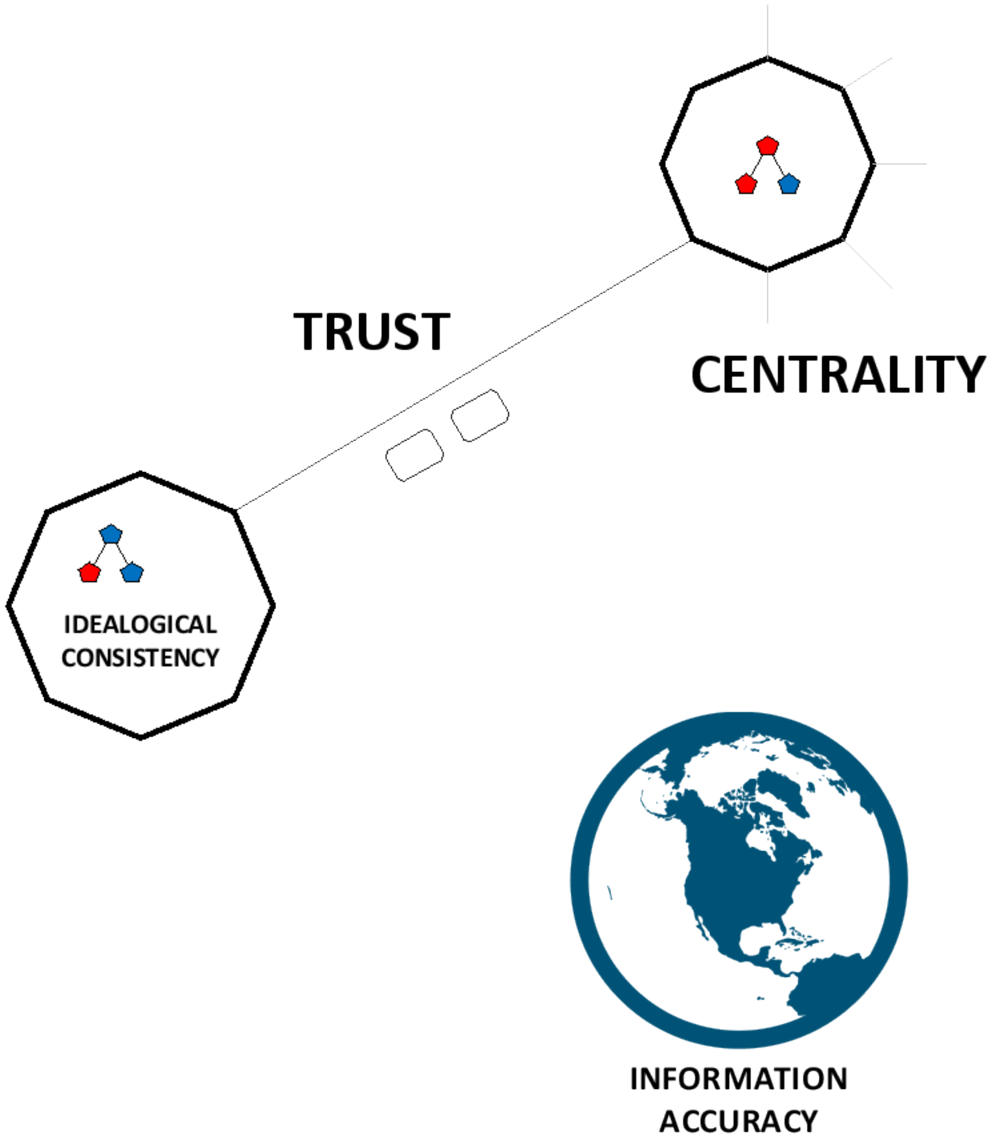
$$P_{m \rightarrow outbox} = Virality_m * Trust_{ij} * Ideological.Consistency_{jm}$$

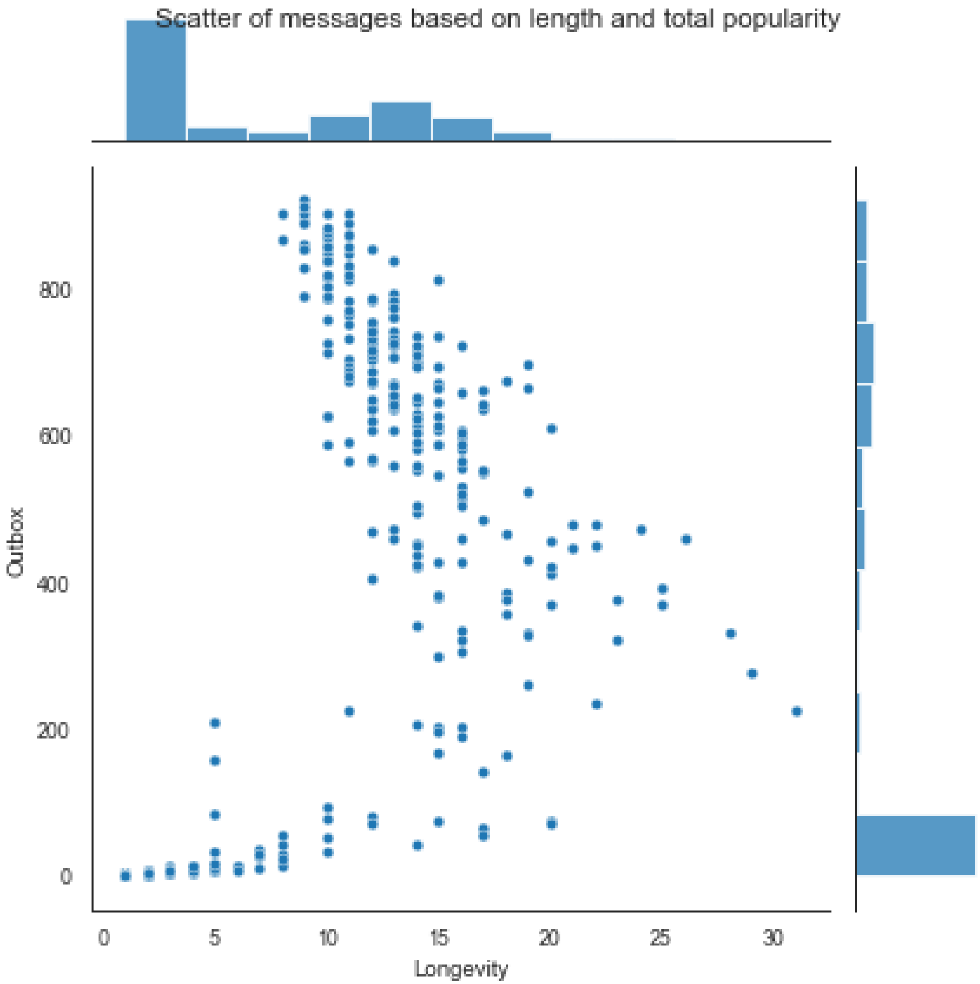
Complex Information Diffusion Model – Information Accuracy

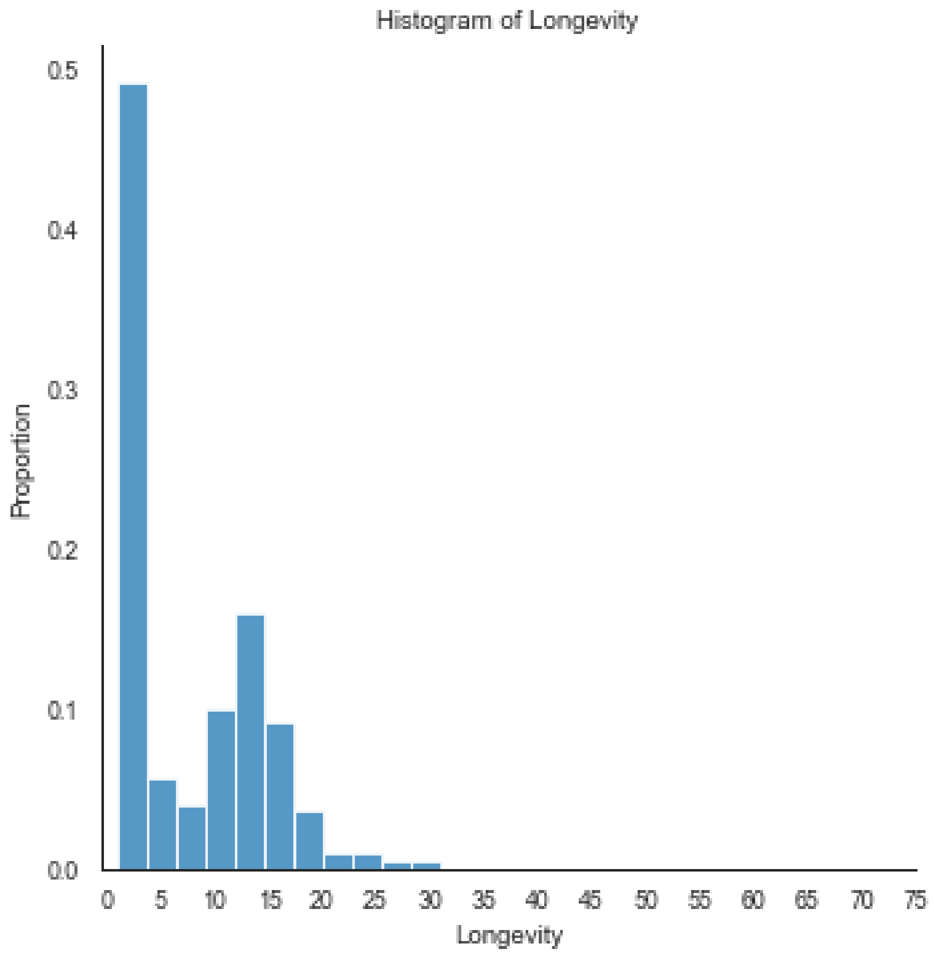
- **Message/receiver characteristic** – the degree to which (receiver's perception of) information in the message conforms with (receiver's perception of) external evidence; true (or perceived true) information is more likely to be adopted and reshared
  - E.g., "vaccines are safe" message paired with evidence of few complications
  - Fairly novel in agent-based models of information diffusion, but interesting because information is modeled as both socially- and externally-supplied
  - One excellent example of its use in ABMs is Lewandowsky et al.'s (2019) model of global warming belief propagation

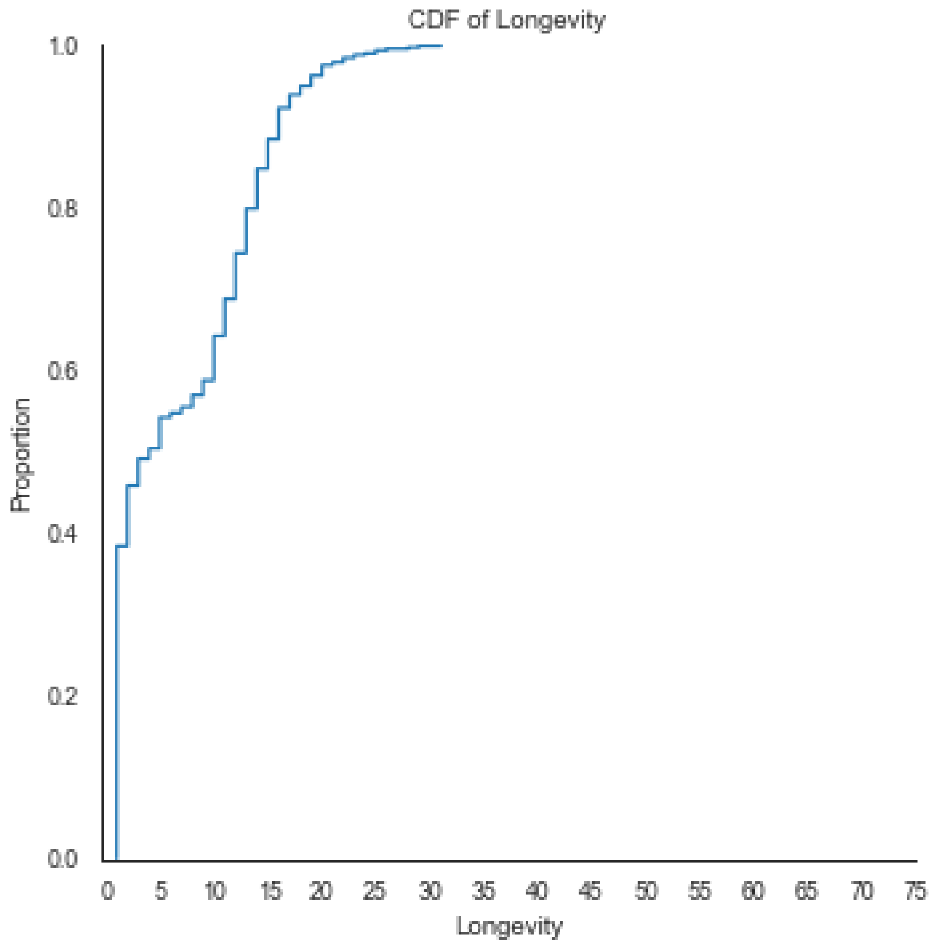
Complex Information Diffusion Model – Information Accuracy

- **In CIDM, information accuracy is operationalized as a filter on read messages** – perceived true information is passed through heuristic processing (trust, virality, ideological consistency), while false information is discarded
- Agents are assigned a knowledge score for each topic (variety of random distributions, {0:1})
- Each message has a random probability of being false (parameterized by topic)
- The probability of detecting that a message is false is given by a sigmoid function tied to knowledge – topic experts are more likely to accurately detect false information than non-experts

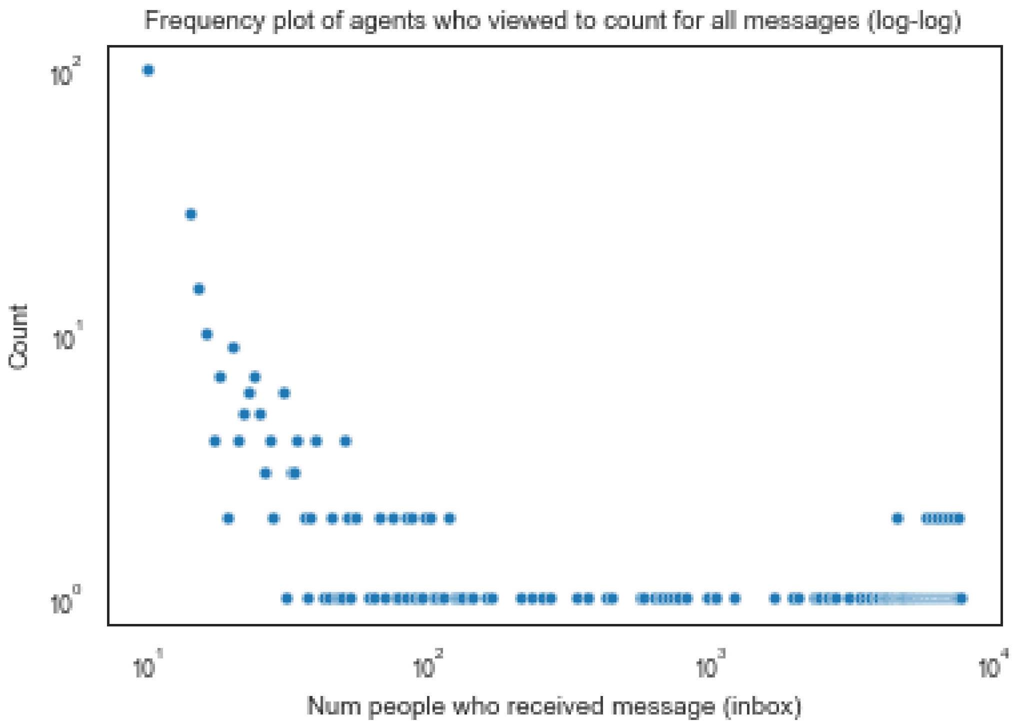

TRUST
CENTRALITY
IDEALOGICAL
CONSISTENCY
INFORMATION
ACCURACY


Scatter of messages based on length and total popularity
Outbox
800
600
400
200
0
0
5
10
15
20
25
30
Longevity


Histogram of Longevity
Proportion
0.0
0.1
0.2
0.3
0.4
0.5
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75
Longevity

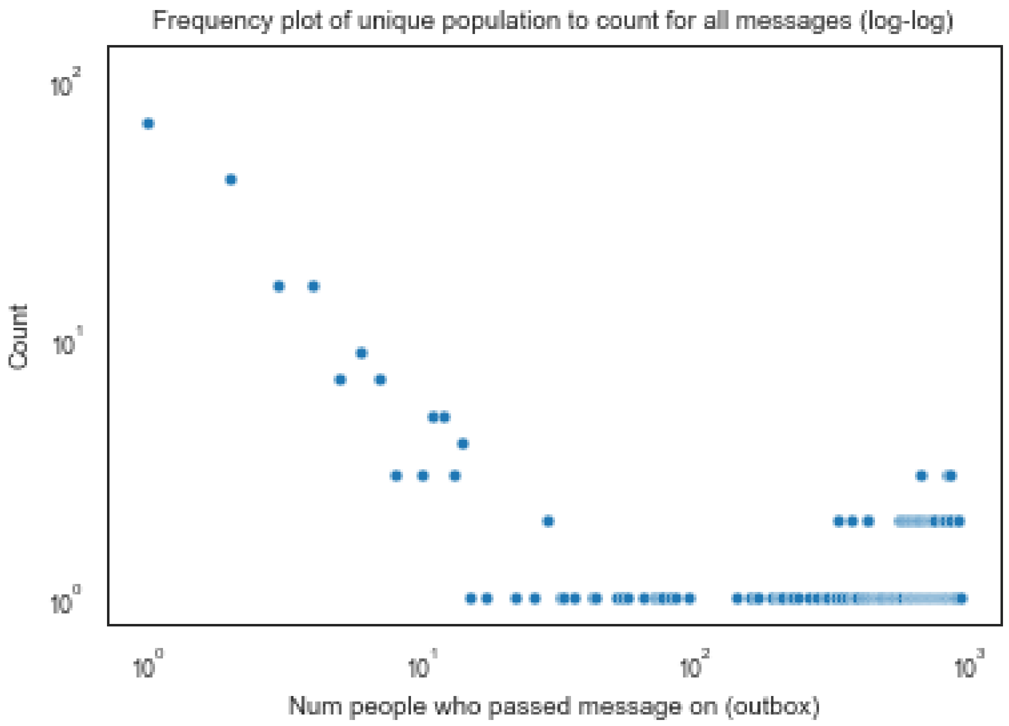

CDF of Longevity
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
Longevity

INBOX



Frequency plot of agents who viewed to count for all messages (log-log)
Count
10^2
10^1
10^0
10^1
10^2
10^3
10^4
Num people who received message (inbox)

OUTBOX



Frequency plot of unique population to count for all messages (log-log)
Count
10^2
10^1
10^0
10^0
10^1
10^2
10^3
Num people who passed message on (outbox)

Conclusions

- Disinformation is a complex problem.
  - National security relevant problems have many of the same issues:
    - Complex interdependencies
    - Lack of data and ground truth.
    - Adversarial setting.
- Social simulations can serve as a testbed:
  - Full ground truth.
  - Can control data bias.
  - Can run experiments and counterfactuals.
  - Can evaluate performance on varied models, parameterizations, etc.

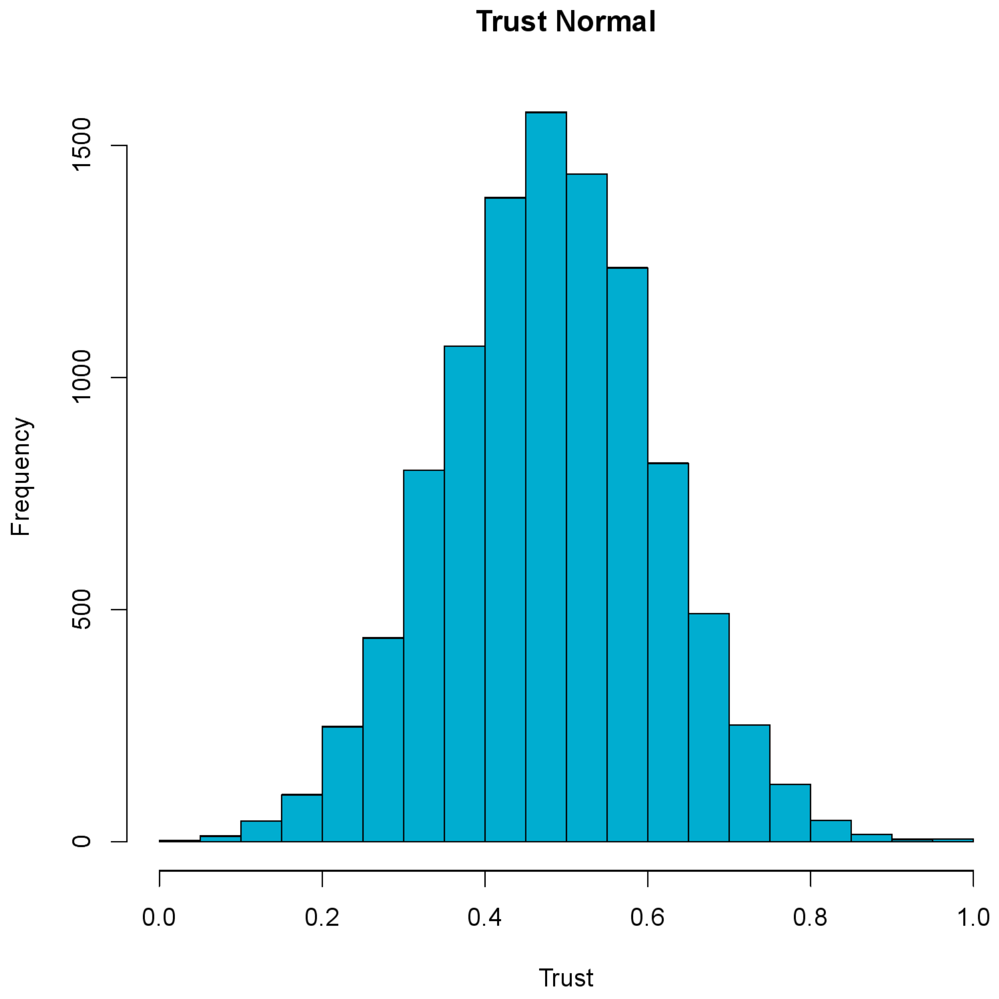
References

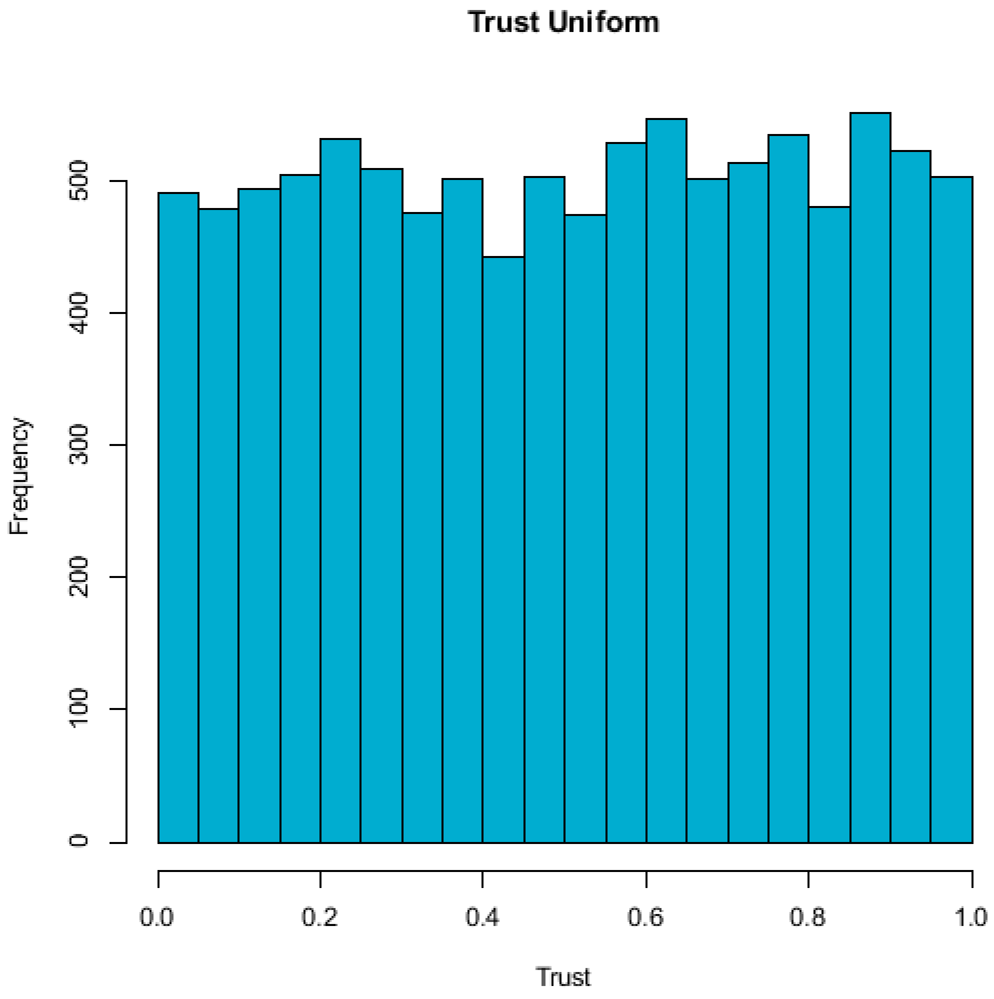
- Bradshaw S, Bailey H, Howard PN (2021) Industrialized Disinformation: 2020 Global Inventory of Organized Social Media Manipulation. Programme on Democracy & Technology, Oxford, UK
- Nguyen CV, Hassner T, Seeger M, Archambeau C (2020) LEEP: a new measure to evaluate transferability of learned representations. In: Proceedings of the 37th International Conference on Machine Learning. JMLR.org, pp 7294–7305
- Pogorelov K, Schroeder DT, Filkuková P, et al (2021) WICO Text: A Labeled Dataset of Conspiracy Theory and 5G-Corona Misinformation Tweets. In: Proceedings of the 2021 Workshop on Open Challenges in Online Social Networks. Association for Computing Machinery, New York, NY, USA, pp 21–25
- Vosoughi S, Roy D, Aral S (2018) The spread of true and false news online. Science 359:1146–1151. https://doi.org/10.1126/science.aap9559
- Zhou F, Xu X, Trajcevski G, Zhang K (2021) A Survey of Information Cascade Analysis: Models, Predictions, and Recent Advances. ACM Comput Surv 54:27:1-27:36. https://doi.org/10.1145/3433000
- Zhuang F, Qi Z, Duan K, et al (2021) A Comprehensive Survey on Transfer Learning. Proceedings of the IEEE 109:43–76. https://doi.org/10.1109/JPROC.2020.3004555

References cont..

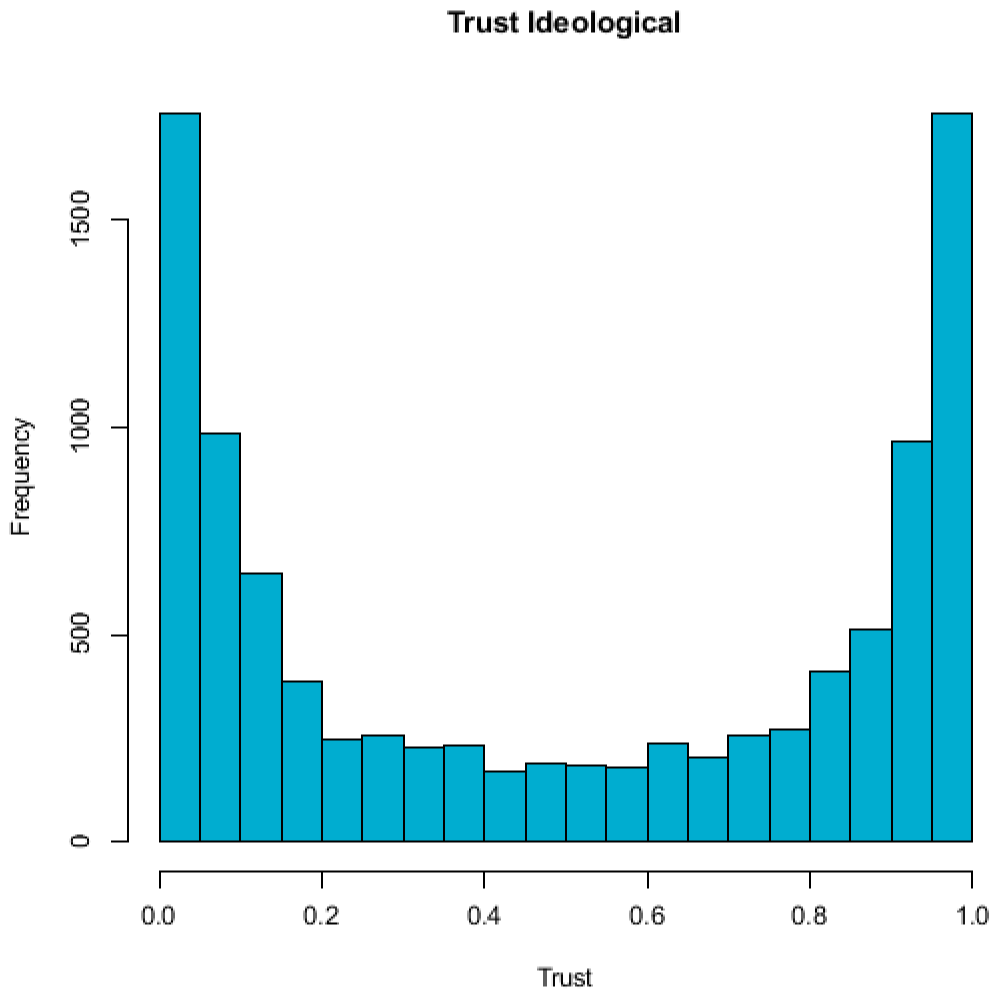
- Barbuto A, Lopolito A, Santeramo FG (2019) Improving diffusion in agriculture: an agent-based model to find the predictors for efficient early adopters. Agricultural and food economics 7(1):1–12
- Berlo DK (1960) The process of communication. New York: Holt, Rinehart, and Winston
- Fan R, Xu K, Zhao J (2018) An agent-based model for emotion contagion and competition in online social media. Physica a: statistical mechanics and its applications 495:245–259
- Festinger L (1962) Cognitive dissonance. Scientific American 207(4):93–106
- Hui C, Goldberg M, Magdon-Ismail M, Wallace WA (2010) Simulating the diffusion of information: An agent-based modeling approach. International Journal of Agent Technologies and Systems (IJATS) 2(3):31–46
- Ibarra H, Andrews SB (1993) Power, social influence, and sense making: Effects of network centrality and proximity on employee perceptions. Administrative science quarterly pp 277–303
- Igarashi T, Kashima Y, Kashima ES, Farsides T, Kim U, Strack F, Werth L, Yuki M (2008) Culture, trust, and social networks. Asian Journal of Social Psychology 11(1):88–101
- Kameda T, Ohtsubo Y, Takezawa M (1997) Centrality in sociocognitive networks and social influence: An illustration in a group decision-making context. Journal of personality and social psychology 73(2):296
- Lakkaraju K (2016) Modeling attitude diffusion and agenda setting: the mama model. Social Network Analysis and Mining 6(1):1–13
- Lewandowsky S, Pilditch TD, Madsen JK, Oreskes N, Risbey JS (2019) Influence and seepage: An evidence-resistant minority can affect public opinion and scientific belief formation. Cognition 188:124–139
- Lotito QF, Zanella D, Casari P (2021) Realistic aspects of simulation models for fake news epidemics over social networks. Future Internet 13(3):76
- Lu Y, Zhang P, Cao Y, Hu Y, Guo L (2014) On the frequency distribution of retweets. Procedia Computer Science 31:747–753
- Schweighofer S, Garcia D, Schweitzer F (2020) An agent-based model of multi-dimensional opinion dynamics and opinion alignment. Chaos: An Interdisciplinary Journal of Nonlinear Science 30(9):093,139
- Sherchan W, Nepal S, Paris C (2013) A survey of trust in social networks. ACM Computing Surveys (CSUR) 45(4):1–33
- Wang H, Zhao J, Li Y, Li C (2015) Network centrality, organizational innovation, and performance: A meta-analysis. Canadian Journal of Administrative Sciences/Revue Canadienne des Sciences de l'Administration 32(3):146–159

Complex Information Diffusion Model – Trust

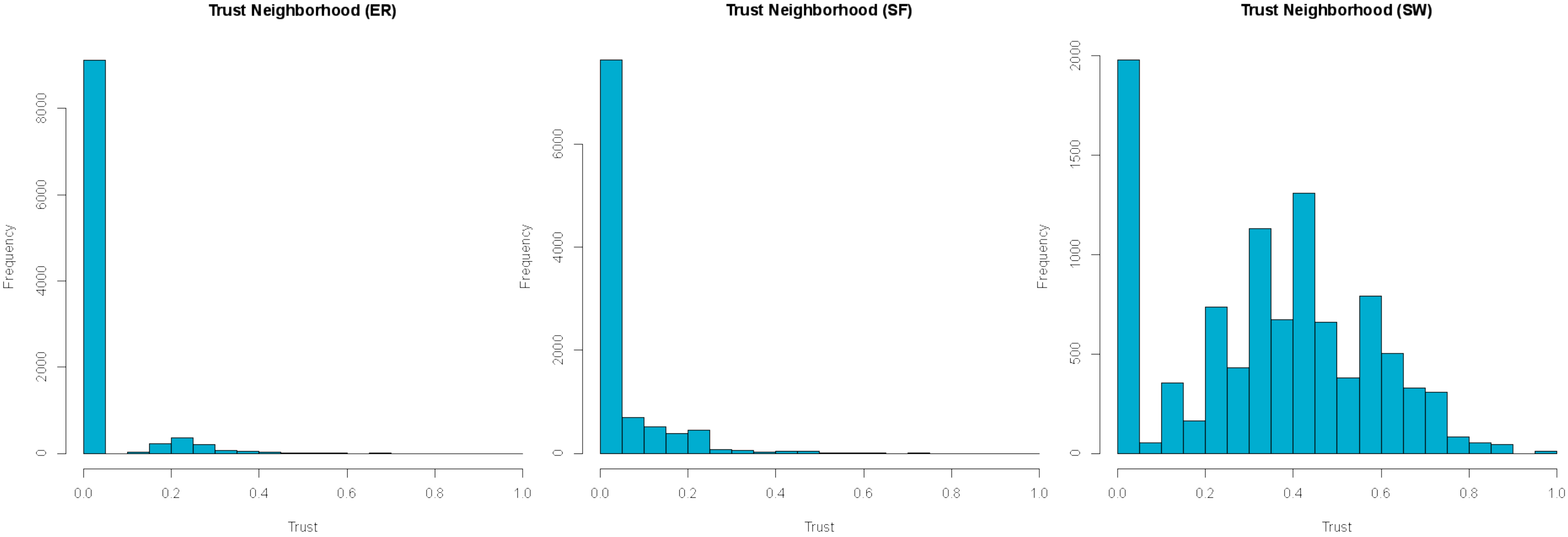




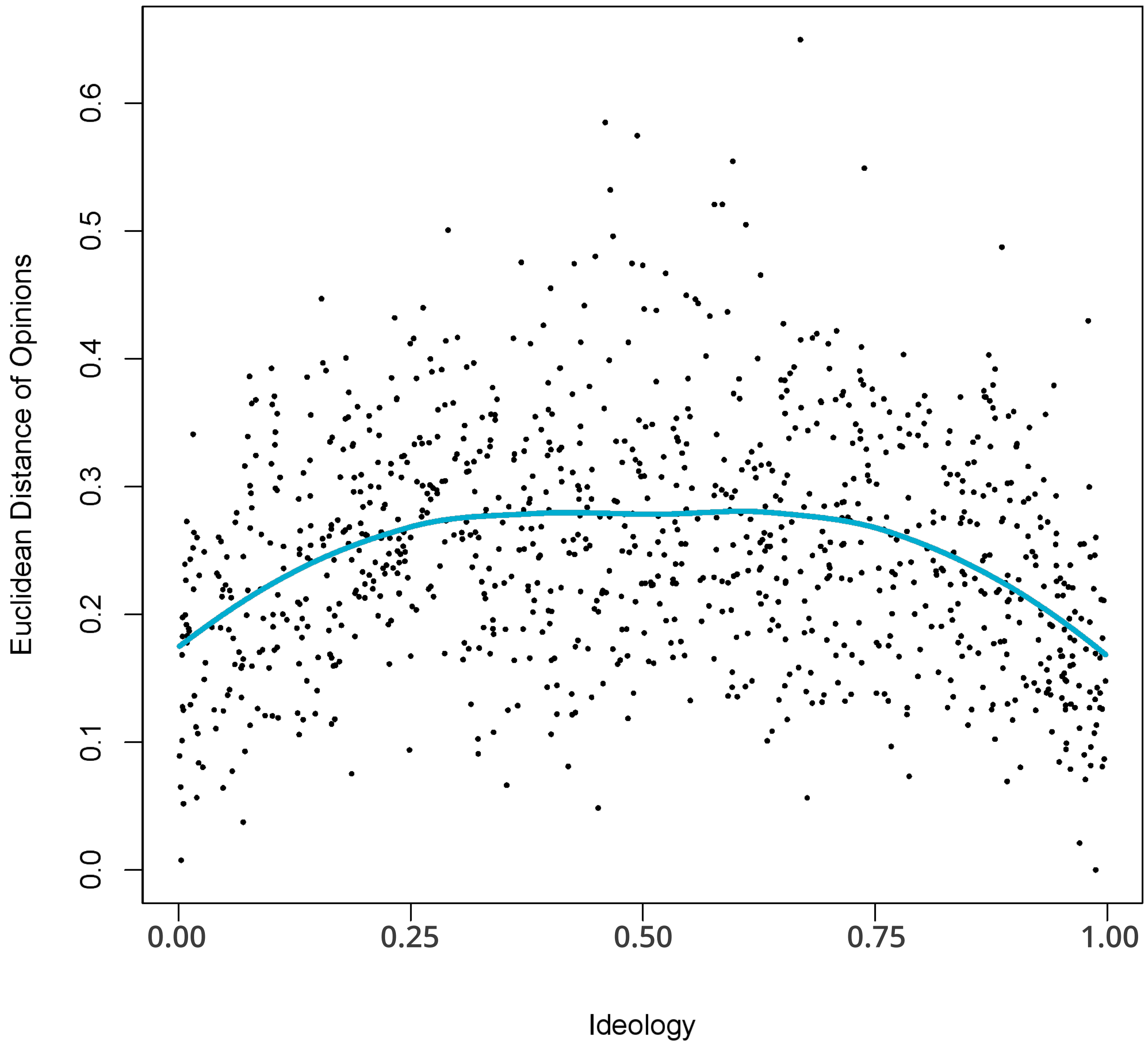
Complex Information Diffusion Model – Trust



Complex Information Diffusion Model – Trust



Complex Information Diffusion Model – Ideological Consistency



Complex Information Diffusion Model – Social Network Centrality

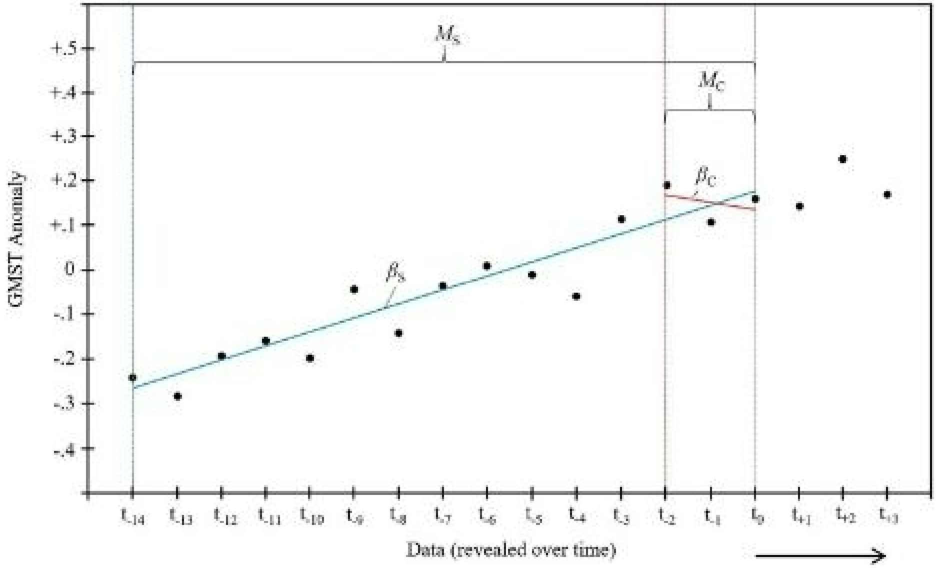
Order	Message	Sender	Centrality
1	102	i	0.83
2	103	i	0.83
3	106	j	0.55
4	102	k	0.52
5	104	l	0.11

$K_i = 3$

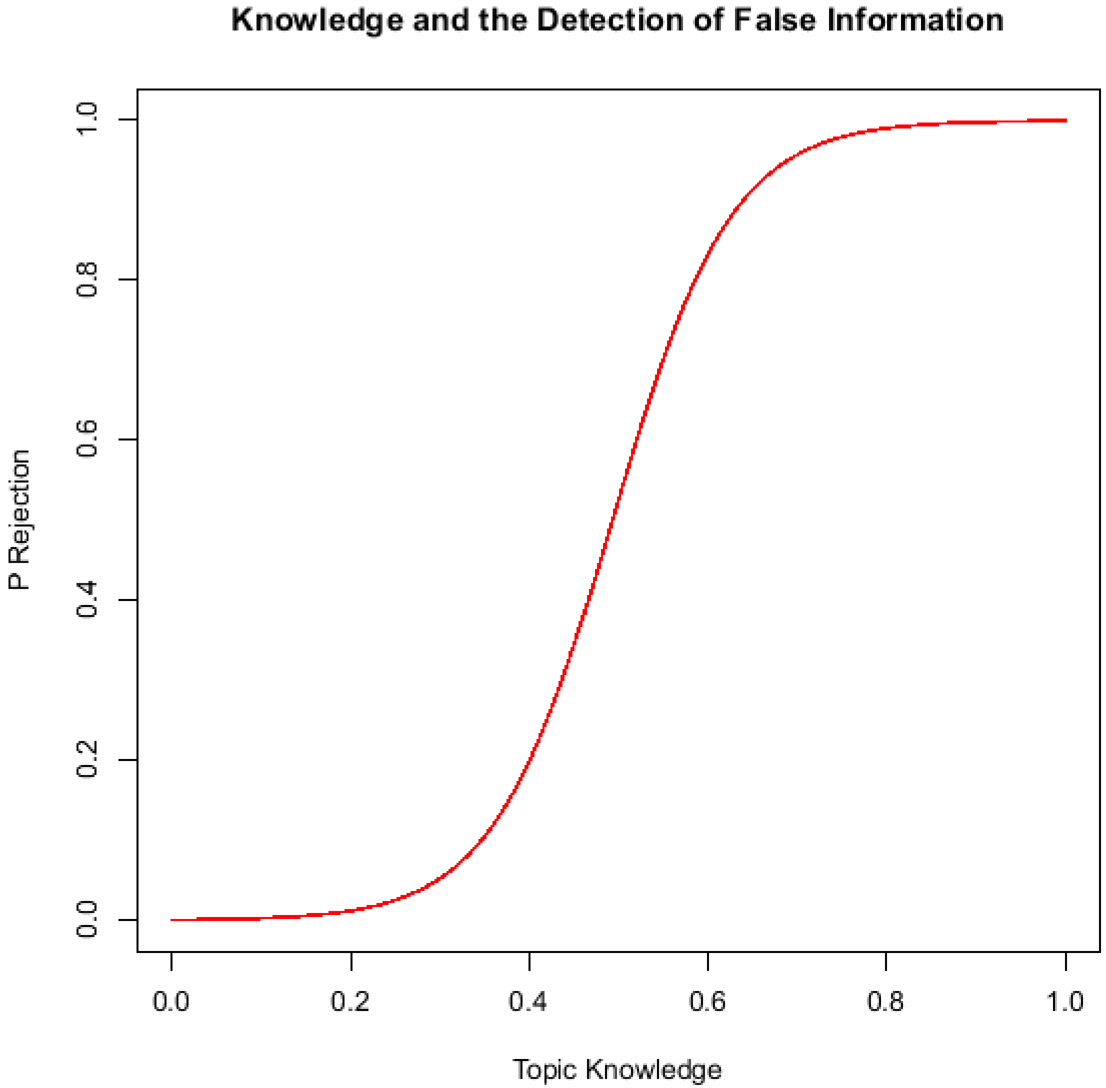
Complex Information Diffusion Model – Information Accuracy

Lewandowsky et al. (2019)

- Three types of agents: scientists, gen. pop., and contrarians
  - Varied the amount of real-world data (last 15-30 years, no data, 3 years) drawn on to form evidence-based opinion on existence of global warming; contrarians apply "skew" (see cognitive consistency)
    - Likelihood ratio drawn from linear regression slope
      - $LR = 10^{\beta - S}$
    - Bayesian belief revision
  - Scientists and contrarians confer within groups
  - They then spread to the general public 5 times per year
- Even small amounts of contrarians drastically reduce overall belief in climate change, both because of skew *and* over-reliance on small amount of data

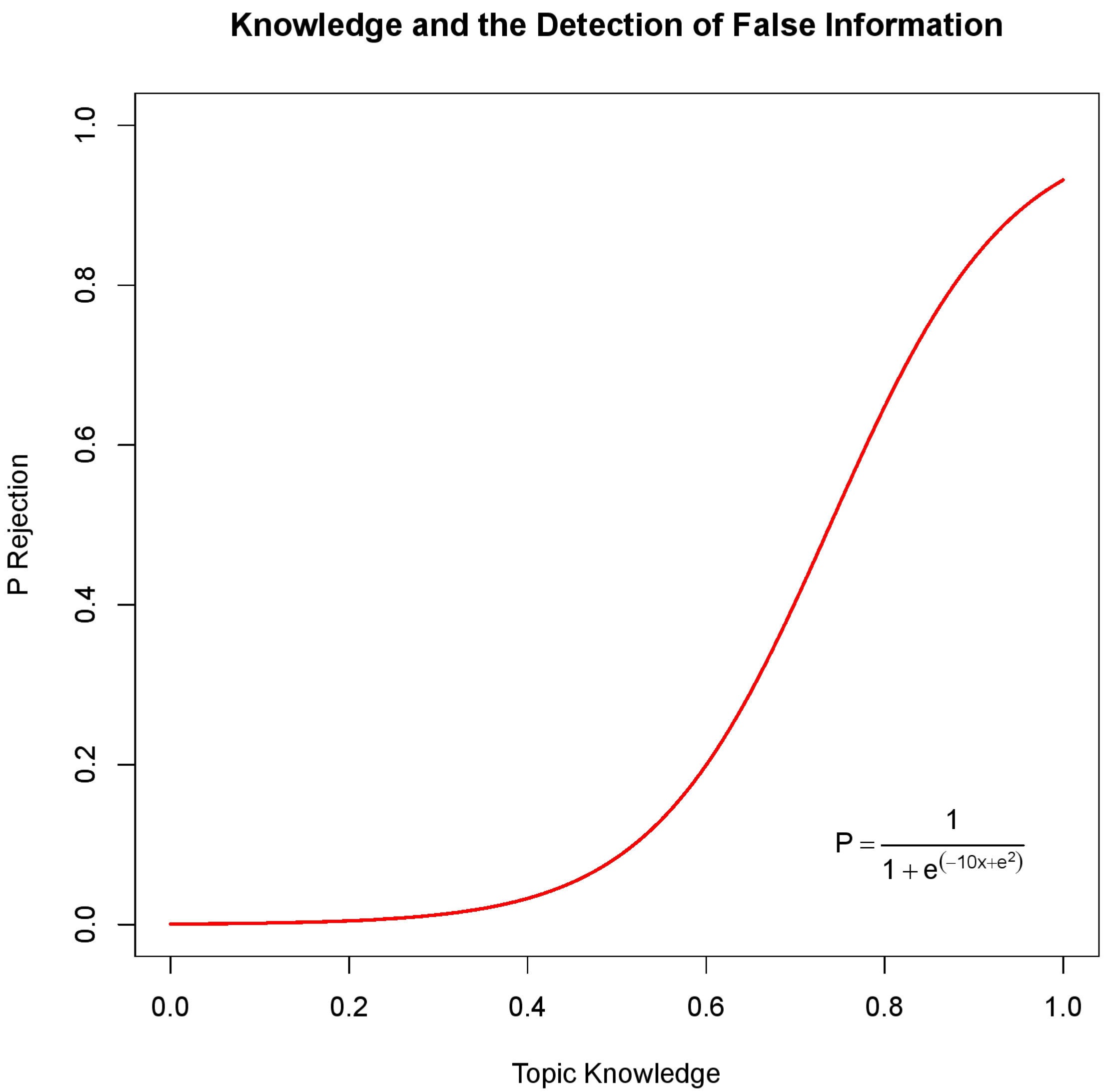


Information Accuracy

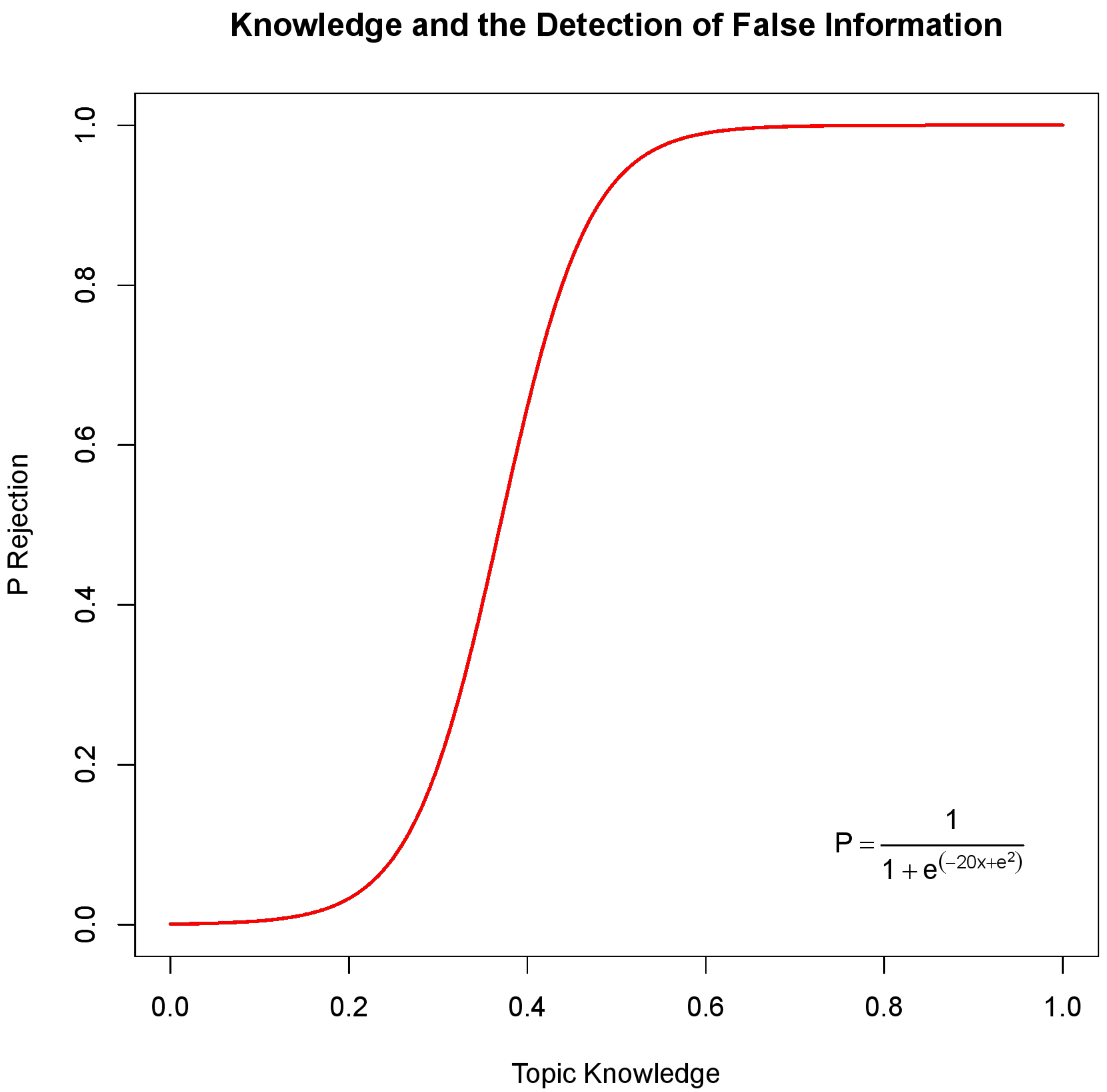


$$P = \frac{1}{1 + e^{(-15x+e^2)}}$$

Information Accuracy



Information Accuracy



Grid Sweep Parameter Settings

- **Number of seeded messages:** 50, 250
- **Number of agents seeded with each new message:** 50
- **Message virality drawn from power distribution with alpha:** 4
- **Number of agents:** 1,000
- **Max number of timesteps:** 100
- **Number of topics:** 3
- **Probability of false message by topic:** (0.1, 0.1, 0.1)
- **Number applied to the false detection sigmoid function by topic:** (4, 4, 4)
- **Add new messages every x ticks:** 5
- **Every x ticks, add mean(SD) messages:** 10(2), 50(10)
- **Network type:** random, scale free, small world
  - **Network density:** 0, 0.008, 0.04
  - **Small world re-wiring probability:** 0, 0.1, 0.5
- **How do distribute trust along all directed edges:** random uniform, 1-mean distance of opinions (ideological homophily)
- **Qi mean(SD) – subjective resend probability:** 1(0.2)
- **Ki mean(SD) – subjective attention limit on inbox:** 5(1), 15(3)
- **How to distribute ideology:** random uniform, random Gaussian (M = 0.3, SD = 0.2)
- **How to distribute topic opinions from ideology:** small random Gaussian (M = ideology, SD = 0.05), large random Gaussian (M = ideology, SD = 0.25)
- **How to distribute topic knowledge:** triangular distribution with mode (0.2, 0.2, 0.2)

*Highlighted parameters were varied in the grid sweep of every unique parameter combination