Social impact theory and its application to social media

Max Pellert

University of Konstanz

Social Media Data Analysis

Outline

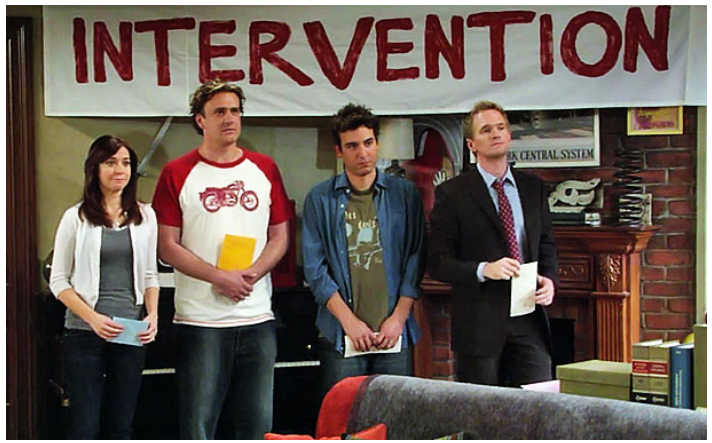
1. Social Impact Theory

2. Social Influence in Online Media

3. Linear Regression in SMDA

What is Social Impact?

> **Social Impact:** Changes in behavior that occur in an individual as a result of presence or actions of other individuals



Examples of behavior: subjective feelings, motives, emotions, thoughts, customs, decisions...

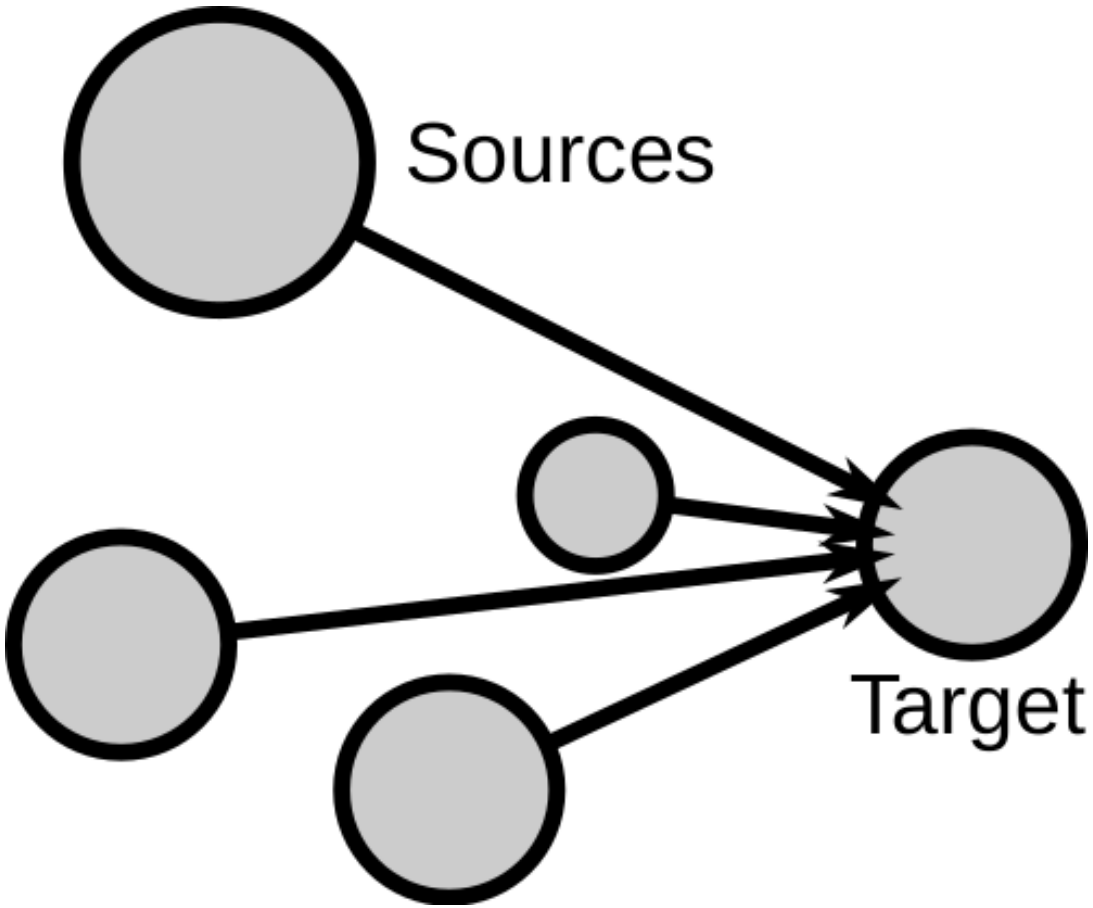
The presence or actions of others can be:

- *real:* the physical presence of others
- *implied:* expected or manipulated presence, e.g. a cardboard policeman
- *imagined:* mental representation of others, e.g. supporters of your team when watching TV sports

Asch's conformity experiments

Elevator experiment in Candid Camera (1962)

Social Impact Theory (Bibb Latané)



Social forces in SIT

In Social Impact Theory, social impact is driven by three forces in the following equation:

$$I = f(S \times i \times N)$$

- I is the magnitude of social impact
- $f()$ is a multiplicative function of three conditions of the impacting situation:
  1. **Strength** S or power of the source(s)
  2. **Immediacy** i or proximity of the source(s)
  3. **Number of sources** N or number of people

Multiplicative effects

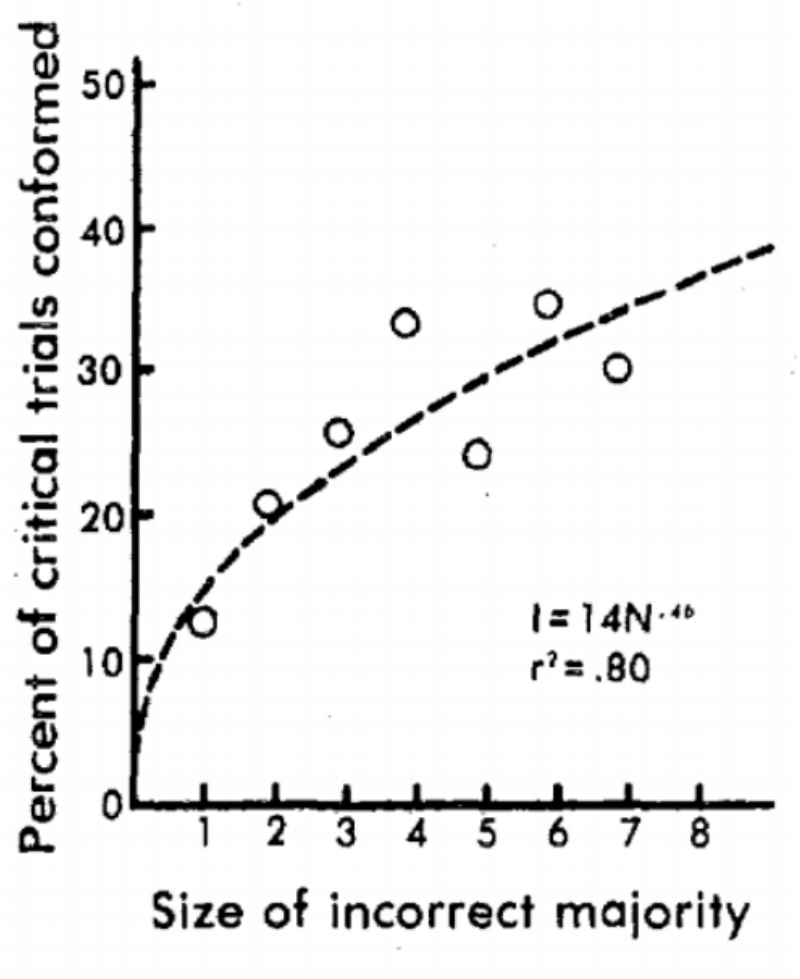
$I = f(S \times i \times N)$ resembles the effect in brightness of a surface illuminated by a number of light sources, their typical strength, and their proximity to the surface.

N: Number of sources

SIT predicts that impact should grow with N. Asch's conformity experiments test this hypothesis, where:

- I: increase in percentage of wrong answers given by students that were experiment subjects
- N: controlled number of confederates

The result: the percentage of wrong answers grows with the amount of sources.



The Psychosocial Law

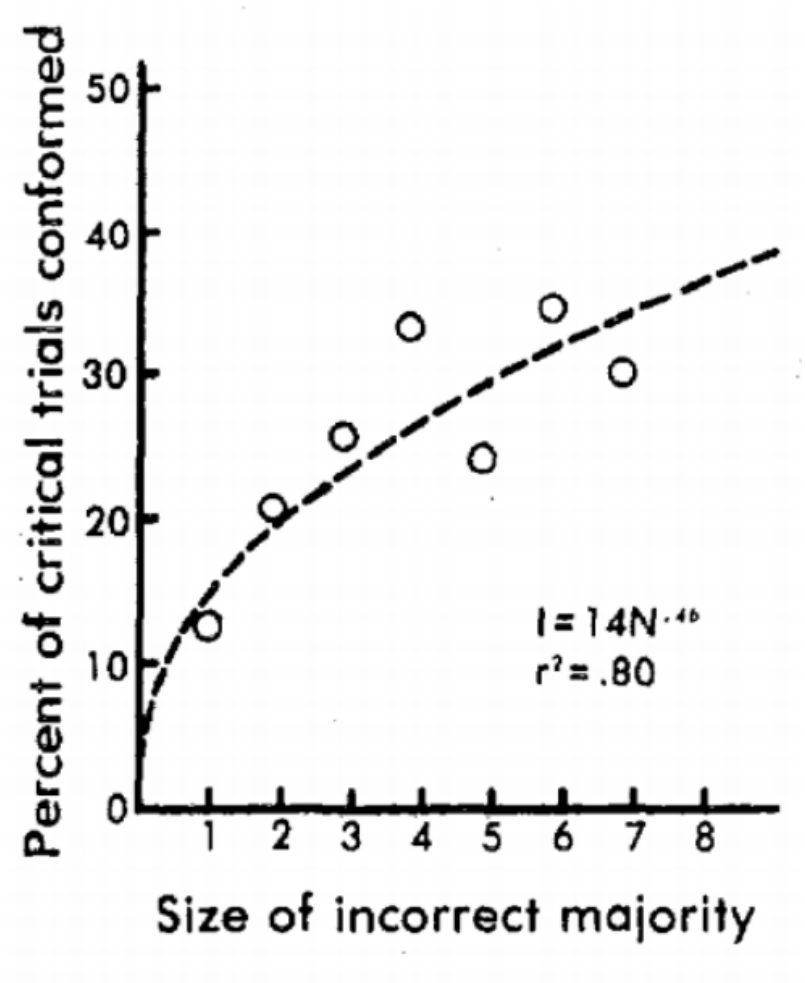
> **The Psychosocial Law:** The extent of social impact grows sublinearly with the number of sources

This can be translated to the equation:

$$I \propto N^t, t < 1$$

This means that the hundredth source has less additional effect than the first (*diminishing returns*).

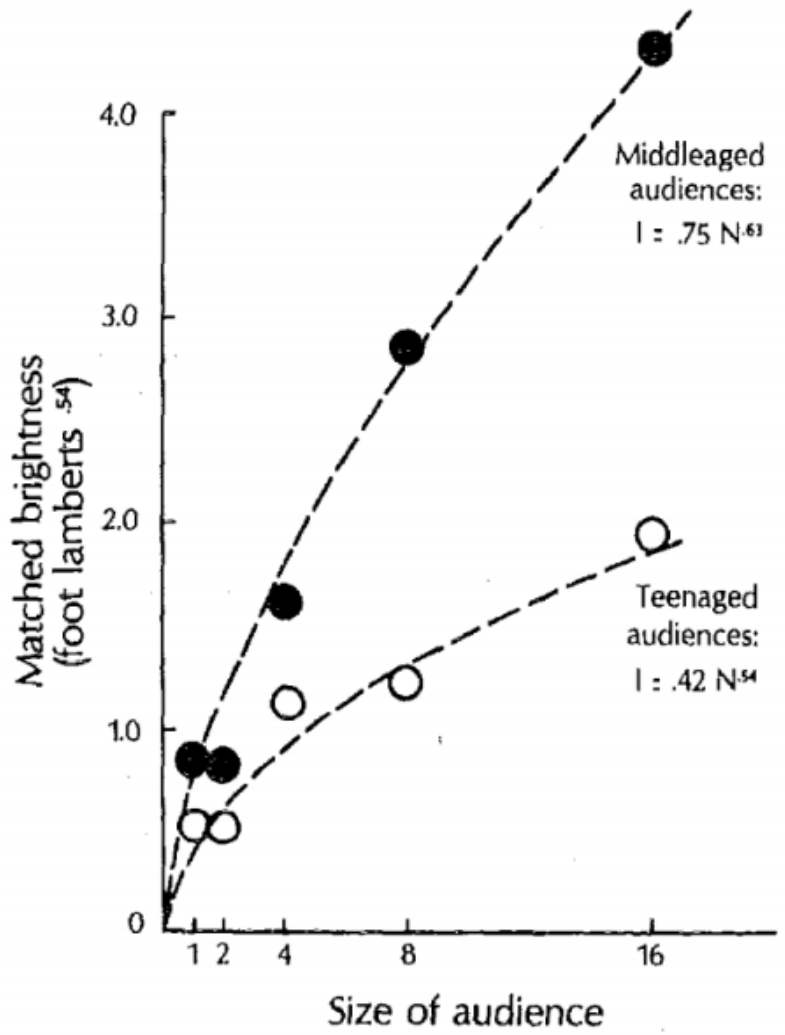
The equation is what is called a power-law with exponent t. In the case of conformity among high school students, t was estimated to be 0.48.



S: Strength of sources

The strength in SIT is the perceived social status, power, wealth, importance, or intensity of the sources. Poem experiment example:

- I: anxiety of the students recorded with a dial they use to measure their own anxiety
- N: number of people in the audience
- S: audience is middle-aged (strong) or teenagers (weak)



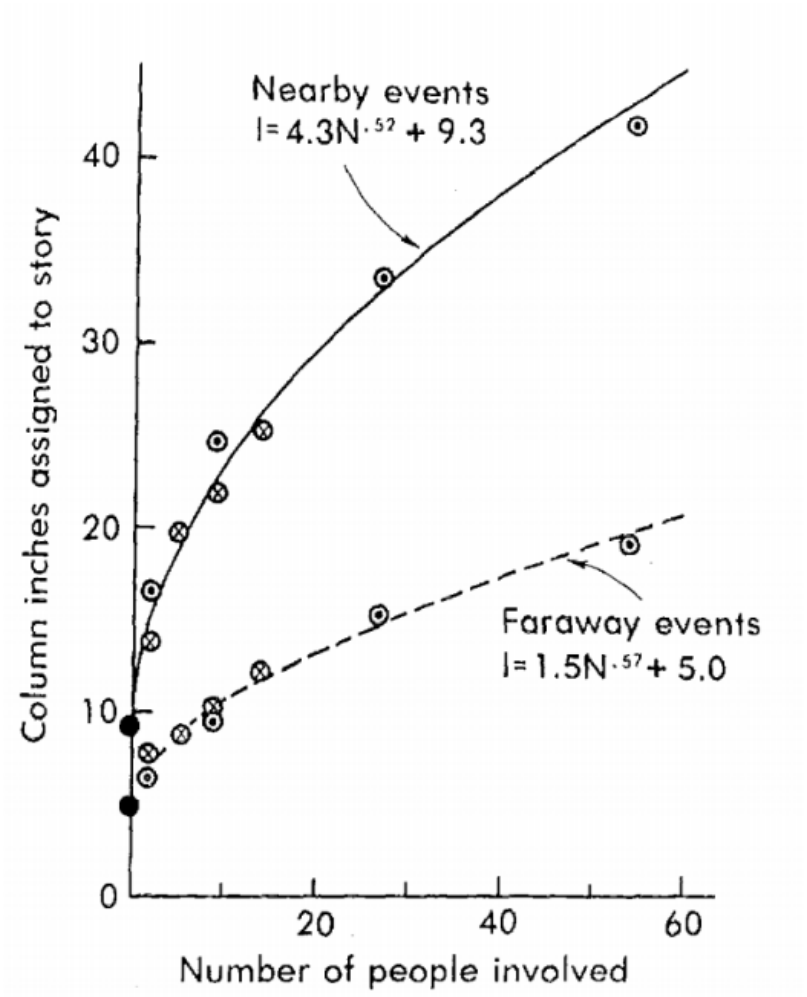
Multiplicative effect: The impact of the number of sources grows faster when they are strong than when they are weak

i: Immediacy of sources

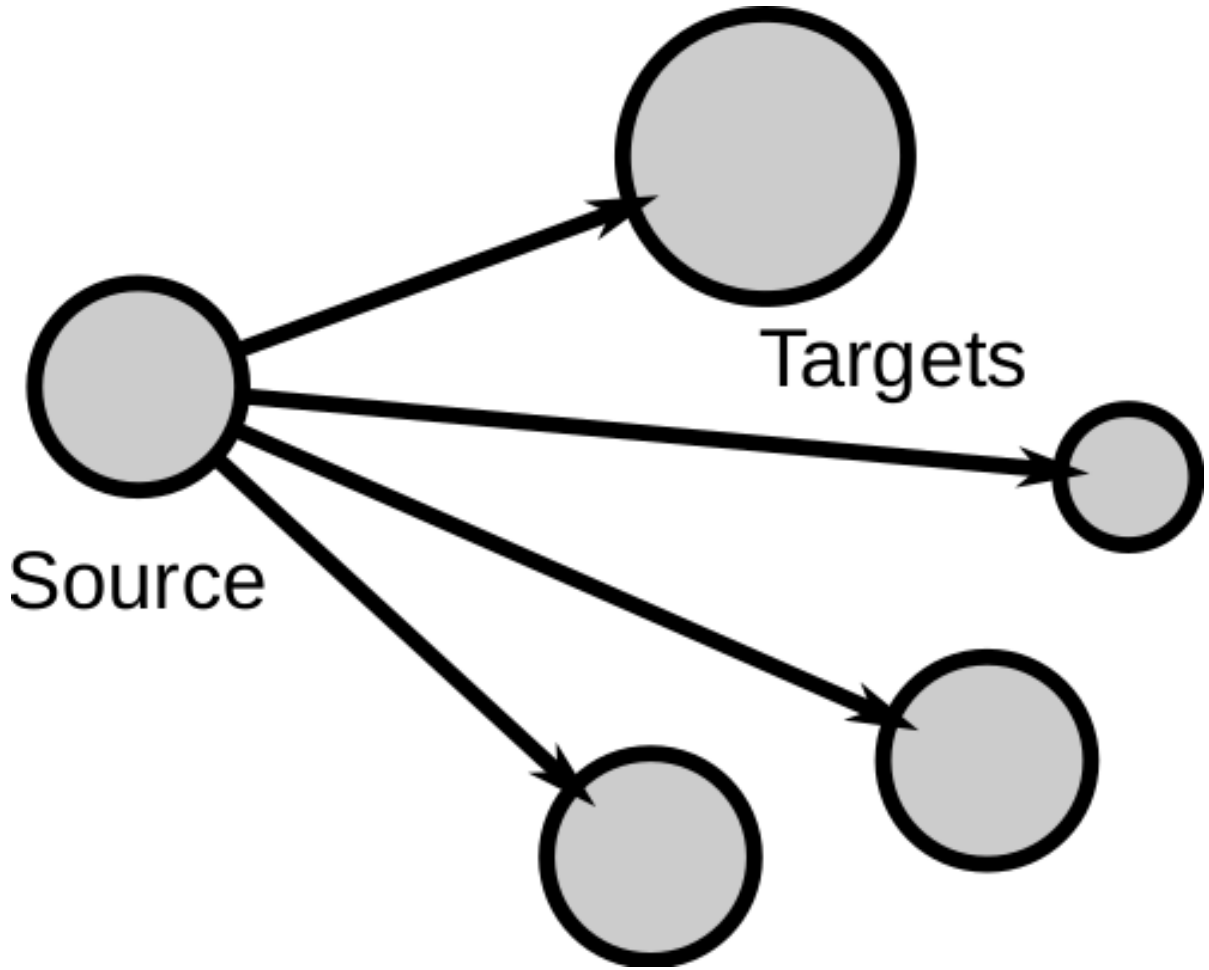
Immediacy is defined as the proximity between the sources and the target of social impact. Immediacy can be spatial, temporal, or social. The effect of immediacy in media bias experiments:

- I: number of lines used to report the news by the students
- N: number of people reported dead in the accident
- i: is the distance to the place of the accident (close vs far conditions)

Results: The growth of impact with N was steeper for the close condition



Division of impact



Division of impact

Social Impact Theory also covers situations with one source but when when targets are not alone. It formulates the impact I on **each** target as:

$$I = f\left(\frac{1}{S \times i \times N}\right)$$

Where the terms are:

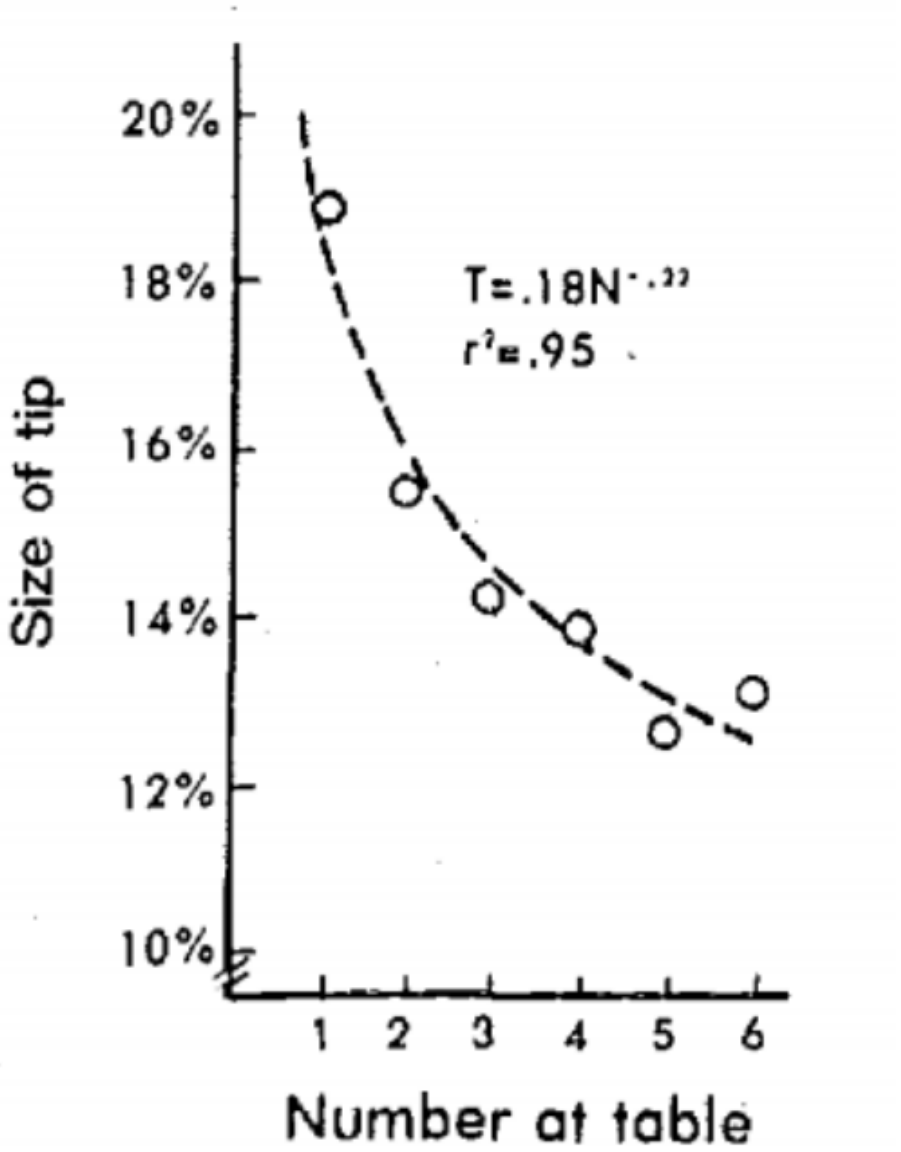
- S strength of the targets: the stronger the targets, the harder to impact each one.
- i immediacy between the targets: the closer or more connected the targets, the harder to impact each one.
- N the more the targets in the group, the harder to impact each one.
- $f()$ is a multiplicative function with negative exponents for the terms.

Division of impact

The most widely studied divisor of impact is group size (N). An observational study finds the effect for the case of restaurant tips:

- I: Percentage of tip (assumed evenly shared among customers)
- N: number of customers at the table

Result: I decreases as N increases. The more people sitting at the same table, the less obliged each one feels to leave a tip. The resulting shape of I as a function of N is well fitted by a negative power of N.

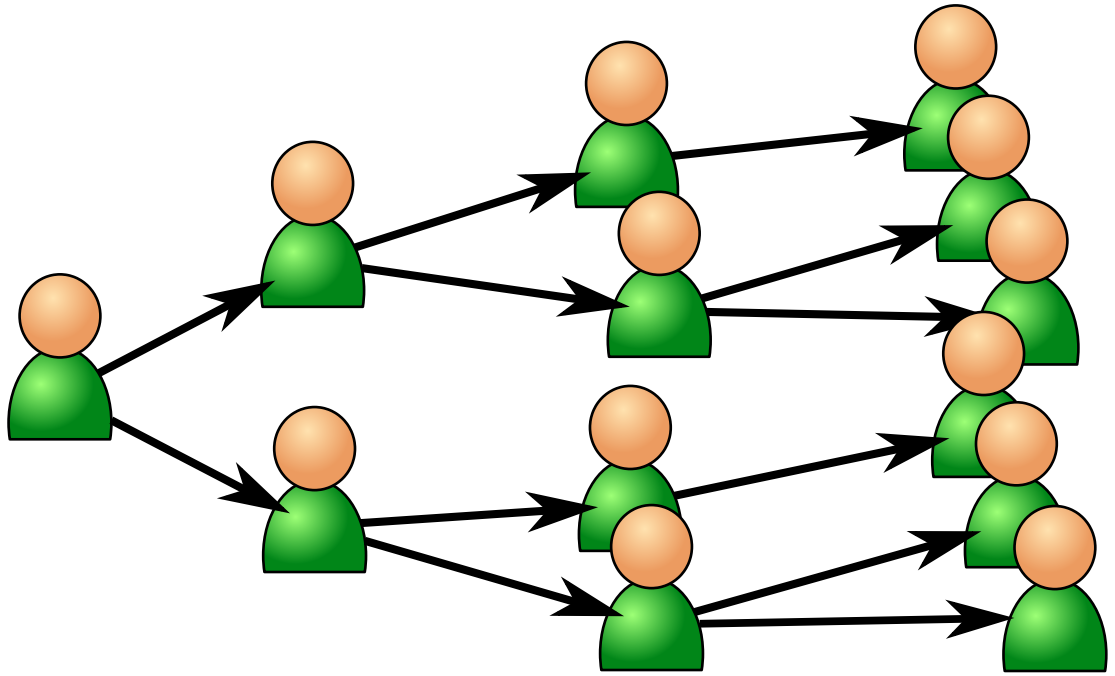


Outline

1. Social Impact Theory
2. *Social Influence in Online Media*
3. Linear Regression in SMDA

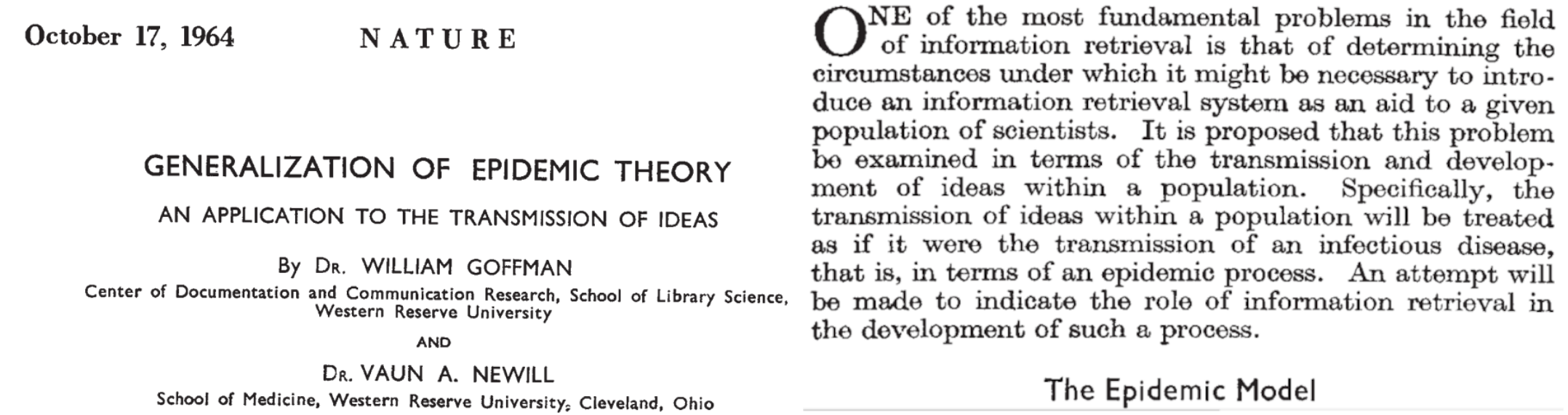
An example of online social influence

The viral analogy of online social sharing



- Humans both transmit diseases and behavior to each other
- The viral analogy illustrates copied behavior and a virus: when some conditions are met, it can reach the whole population
- This analogy is the basis of viral marketing: what makes content spread?

The origin of the viral analogy



October 17, 1964 NATURE

GENERALIZATION OF EPIDEMIC THEORY

AN APPLICATION TO THE TRANSMISSION OF IDEAS

By Dr. WILLIAM GOFFMAN
Center of Documentation and Communication Research, School of Library Science, Western Reserve University
AND
Dr. VAUN A. NEWILL
School of Medicine, Western Reserve University, Cleveland, Ohio

ONE of the most fundamental problems in the field of information retrieval is that of determining the circumstances under which it might be necessary to introduce an information retrieval system as an aid to a given population of scientists. It is proposed that this problem be examined in terms of the transmission and development of ideas within a population. Specifically, the transmission of ideas within a population will be treated as if it were the transmission of an infectious disease, that is, in terms of an epidemic process. An attempt will be made to indicate the role of information retrieval in the development of such a process.

The Epidemic Model

Generalization of Epidemic Theory: An Application to the Transmission of Ideas. Goffman & Newill, Nature (1964)

Formalizing the analogy

Table 1. ANALOGY BETWEEN INFECTIOUS DISEASE AND INTELLECTUAL EPIDEMICS

Elements of the epidemic process	Elements interpreted in terms of Infectious disease epidemic	Intellectual epidemic
Host		
Agent	Infectious material	Idea
Infective	Case of disease	Author of paper
Susceptible	Person who will be infected given effective contact	Reader of paper who will be infected given effective contact
Removal	Death or immunity	Death or loss of interest
Vector		
Agent	Infectious material (as for host)	Idea (as for host)
Infective	Vector harbouring the agent	Paper containing useful ideas
Susceptible	Vector not harbouring the agent	All papers containing potentially useful ideas
Removal	Death	Deletion or loss

Formalizing the analogy

Element	Infectious disease epidemic	Intellectual epidemic	Online epidemic
Agent	Infectious material	Idea in paper	Online post
Infective	Case of disease	Author of paper	User posting social media content
Susceptible	Person infected in face to face contact	Reader influenced by paper	User liking or resharing content
Removal	Death or immunity	Death or loss of interest	User forgetting or changing attention

Simple versus complex contagion

The spreading analogy implies **simple contagion**: each exposure has a probability of behavior adoption independent of anything else

Complex contagion: multiple exposures might be necessary to enable any contagion, especially for risky behavior

Social Impact Theory predicts that:

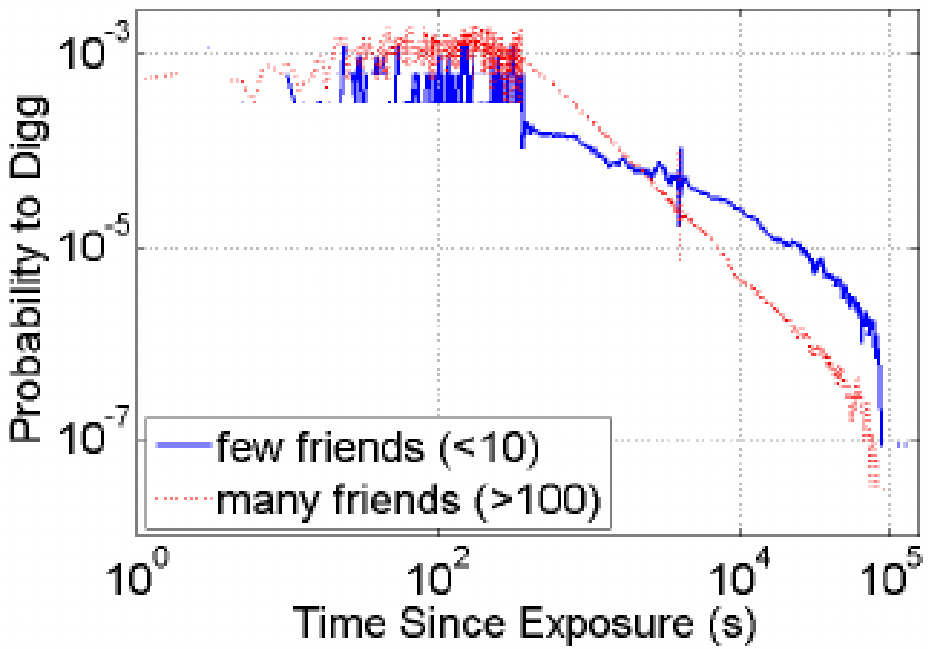
- the number of sources will matter in a sublinear way
- immediacy will have a positive effect in influence
- not all sources are equal in strength when influencing others

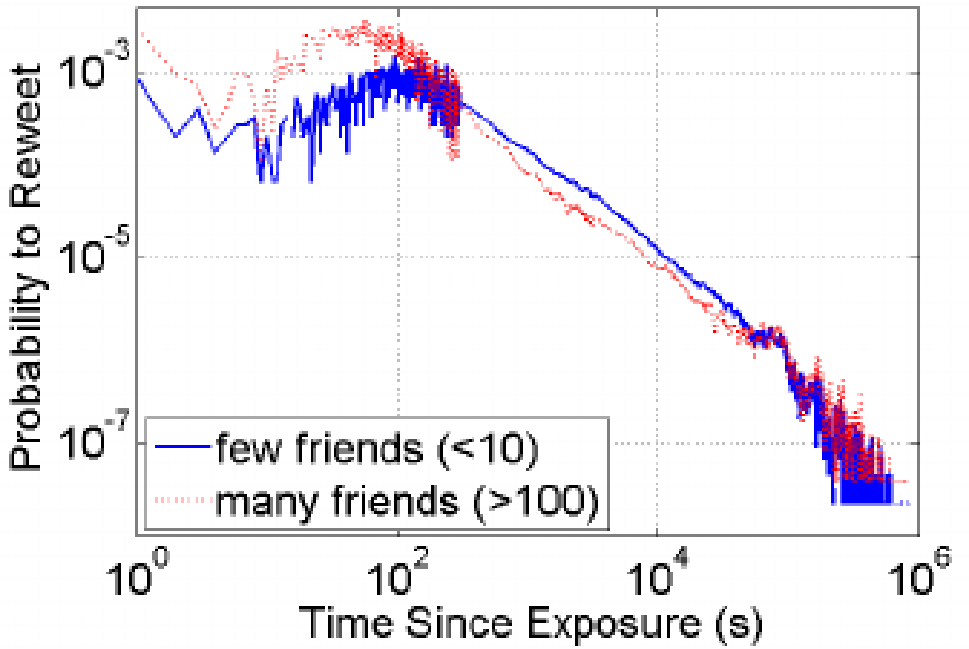
Complex Contagions and the Weakness of Long Ties. Centola & Michael Macy. American Journal of Sociology (2007)

Revising the viral analogy in social media

- Social impact on Twitter: Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. Romero, Meeder & Kleinberg, WWW conference (2011).
- Review including Digg data analysis: Information Is Not a Virus, and Other Consequences of Human Cognitive Limits, Kristina Lerman, Future Internet (2016).

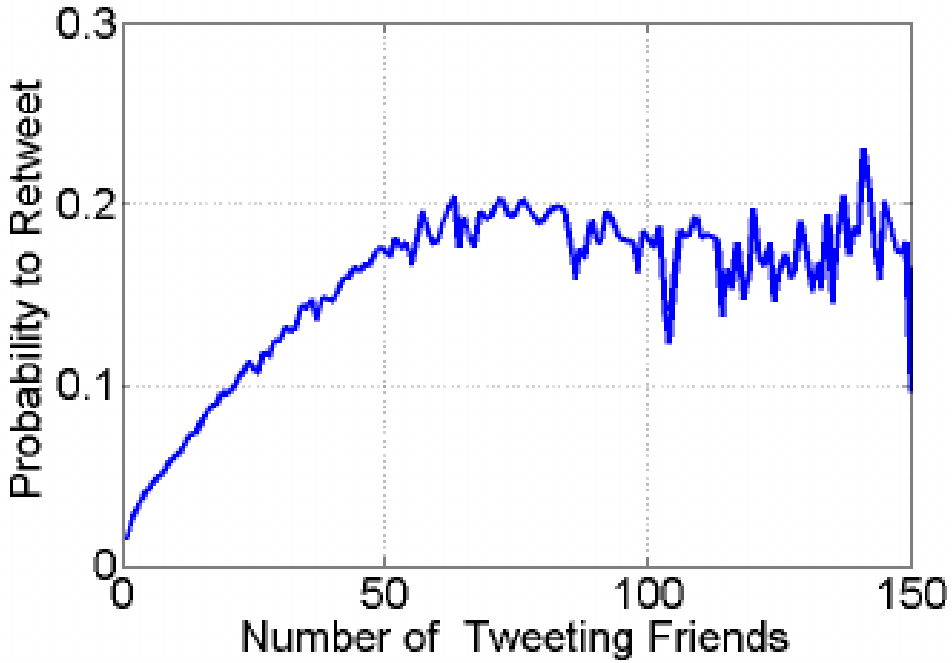
Effect of immediacy

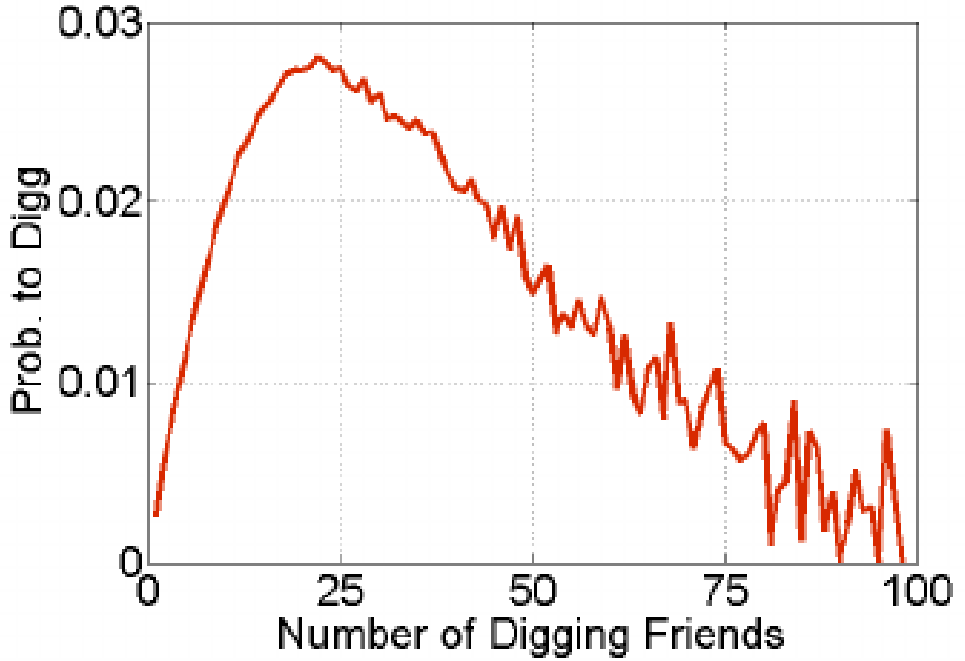




- Probability to adopt behavior is not constant over time: temporal immediacy
- Probability to adopt behavior decreases very fast with time since exposure

Limits to the psychosocial law online



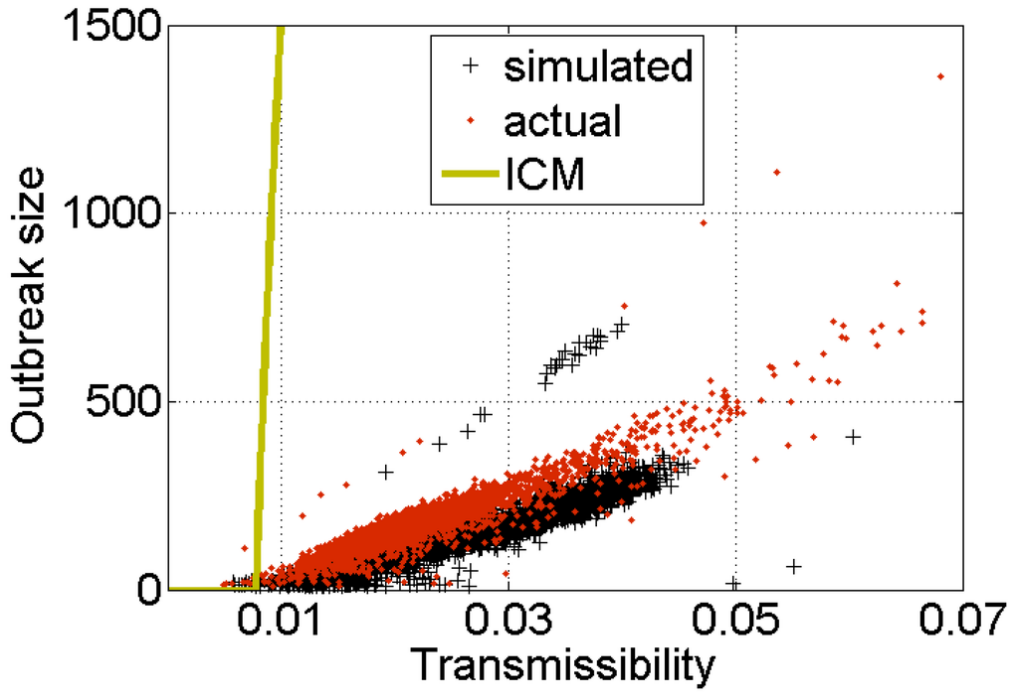


- Probability of adoption versus number of sources
- Growth in impact saturates: not a case of simple contagion
- Growth in impact stops at some point or even reverses. Effect of information overload not hypothesized by Social Impact Theory.

When spreading is not so viral

Independent Cascade Model (ICM)

- Simulation on social network without considering number of exposures and immediacy
- Comparison versus empirical size of cascades (digg number) and simulations including SIT
- ICM wildly overestimates cascades sizes: things are not "so viral"



Information Is Not a Virus, and Other Consequences of Human Cognitive Limits, Kristina Lerman, Future Internet (2016)

Division of Impact in Social Media

In exercise 2, you will test the division of impact hypothesis:

- On a Twitter US politicians dataset
  - Aggregating the average number of retweets versus number of followers
- On your own sample of YouTube channels through the API
  - Measuring mean number of views versus number of subscribers
- Fitting a model to test if impact is sublinear with audience size

Outline

1. Social Impact Theory
2. Social Influence in Online Media
3. *Linear Regression in SMDA*

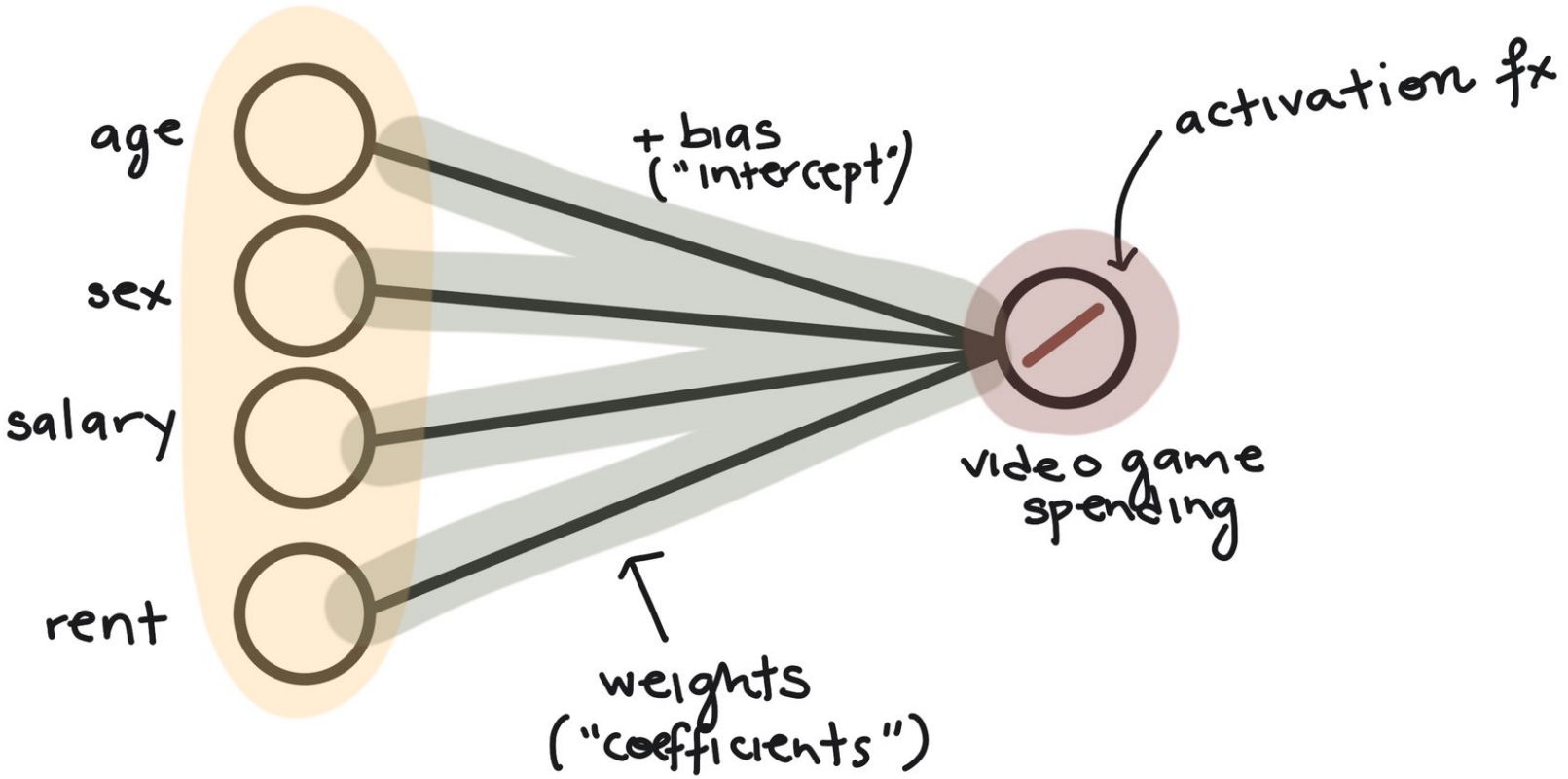
Linear Regression

Regression models formalize an equation in which one numeric variable Y is formulated as a linear function of other variables X_1, X_2, X_3, etc:

$$Y = a + b_1 X_1 + b_2 X_2 + b_3 X_3 \ldots + \epsilon$$

- Y is called the dependent variable
- X_1, X_2, X_3, etc are called independent variables
- a is the intercept, which measures the expected value of Y that does not depend on the dependent variables
- b_1, b_2, b_3, etc are called the slopes or the coefficients
- ϵ are the residuals, the errors of the equation in the data

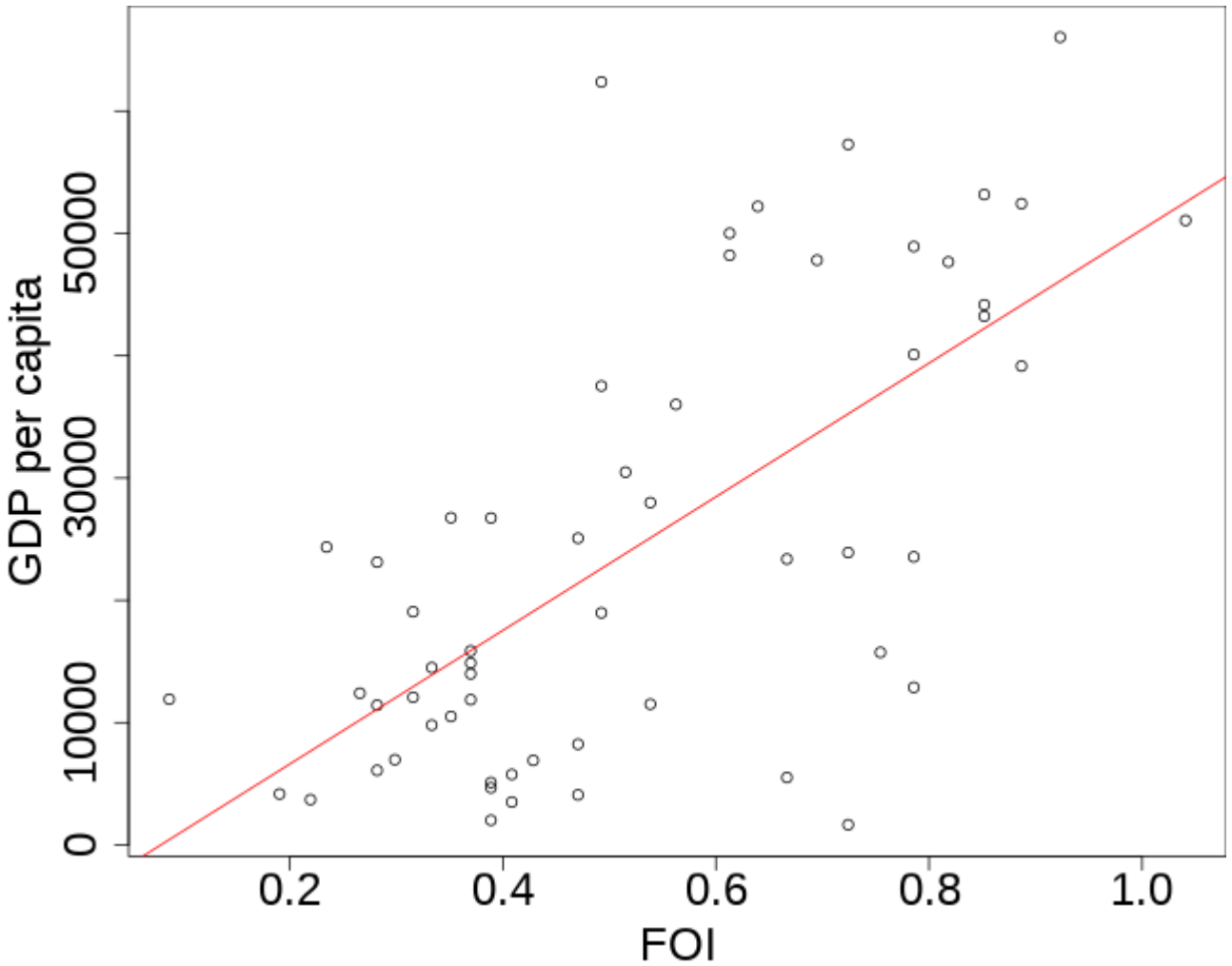
LINEAR REGRESSION
(as a neural network)



$\text{LOSS}: \Sigma(x_i - \tilde{x})^2$

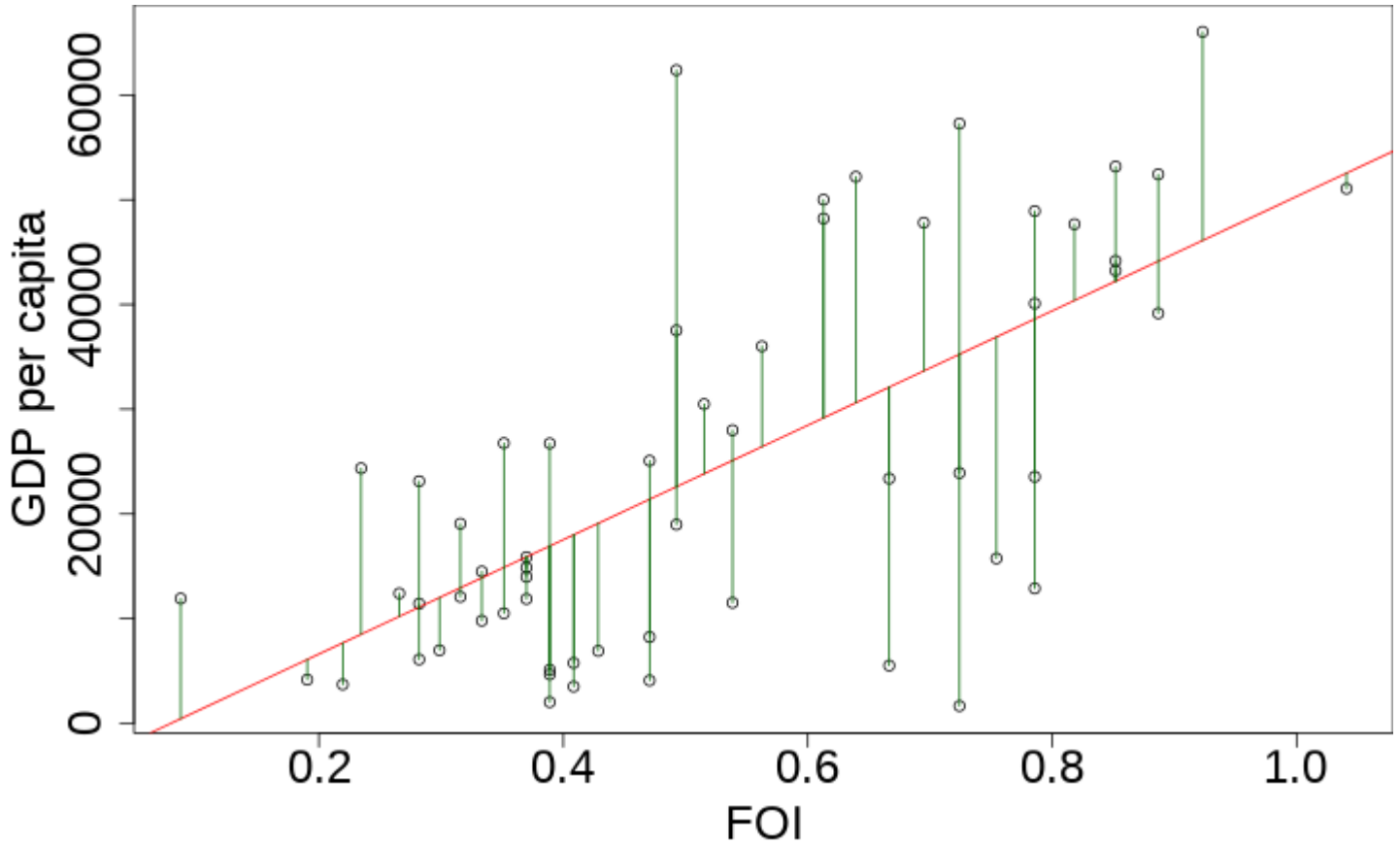
@CHELSEAPARLETT

Example: FOI vs GDP



Regression residuals

Residuals (ϵ) are the differences in between the empirical values Y_i and their fitted values $\hat{Y}_i$.



Ordinary Least Squares (OLS)

Fitting a regression model is the task of finding the values of the coefficients (a, b_1, b_2, etc) in a way that reduce a way to aggregate the residuals of the model. One approach is called Residual Sum of Squares (RSS), which aggregates residuals as:

$$RSS = \sum_i (\hat{Y}_i - Y_i)^2$$

The Ordinary Least Squares method (OLS) looks for the values of coefficients that minimize the RSS. This way, you can think about the OLS result as the line that minimizes the sum of squared lengths of the vertical lines in the figure above.

Goodness of fit

A way to measure the quality of a model fit this is to calculate the proportion of variance of the dependent variable ($V[Y]$) that is explained by the model. We can do this by comparing the variance of residuals ($V[\epsilon]$) to the variance of Y.

This is captured by the coefficient of determination, also known as R^2:

$$R^2 = 1 - \frac{V[\epsilon]}{V[Y]}$$

For our model example, the R^2 is 0.4432583

Model likelihood

- Likelihood $\hat{L}$, probability of the observed data y given the model M and our estimation of its parameters $\hat{\Theta}$:

$$\hat{L} = p(y|\hat{\Theta}, M)$$

- It is calculated as the product of the likehood of each observation

$$\hat{L} = p(y_1|\hat{\Theta}, M) * p(y_2|\hat{\Theta}, M) \dots p(y_N|\hat{\Theta}, M) = p(y|\hat{\Theta}, M)$$

The case of linear regression

Our formulation of the regression model:

$$Y = a + b_1 X_1 + b_2 X_2 + b_3 X_3 \ldots + \epsilon$$

With the assumption that residuals are unbiased, independent, and normal:

$$\epsilon \sim N(0, \sigma^2)$$

Can be reformulated as:

$$Y \sim N(a + b_1 X_1 + b_2 X_2 + b_3 X_3 \ldots, \sigma^2)$$

Where our parameter vector is: $\Theta = [a, b_1, b_2, b_3 \ldots]$

The case of linear regression

We can write it as a conditional probability using the Gaussian probability density function:

$$f(Y|X;\Theta,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp(\frac{-(Y-(a+b_1X_1+b_2X_2+b_3X_3\dots))^2}{2\sigma^2})$$

And this way calculate the likelihood function:

$$L(\Theta,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \prod_{(y,x_1,x_2,x_3\dots)} exp(\frac{-(y-(a+b_1x_1+b_2x_2+b_3x_3\dots))^2}{2\sigma^2})$$

Maximizing the likelihood for a given data is a way of fitting model parameters and comparing models

Bayesian and Akaike Information Criteria

- Likelihood functions do not consider the number of parameters of a model
  - Risk of overfitting: it is easier to get a higher likelihood L with more parameters
- Information criteria correct for the number of parameters k and sample size N so we can compare models with different numbers of parameters:
  - Bayesian Information Criterion:

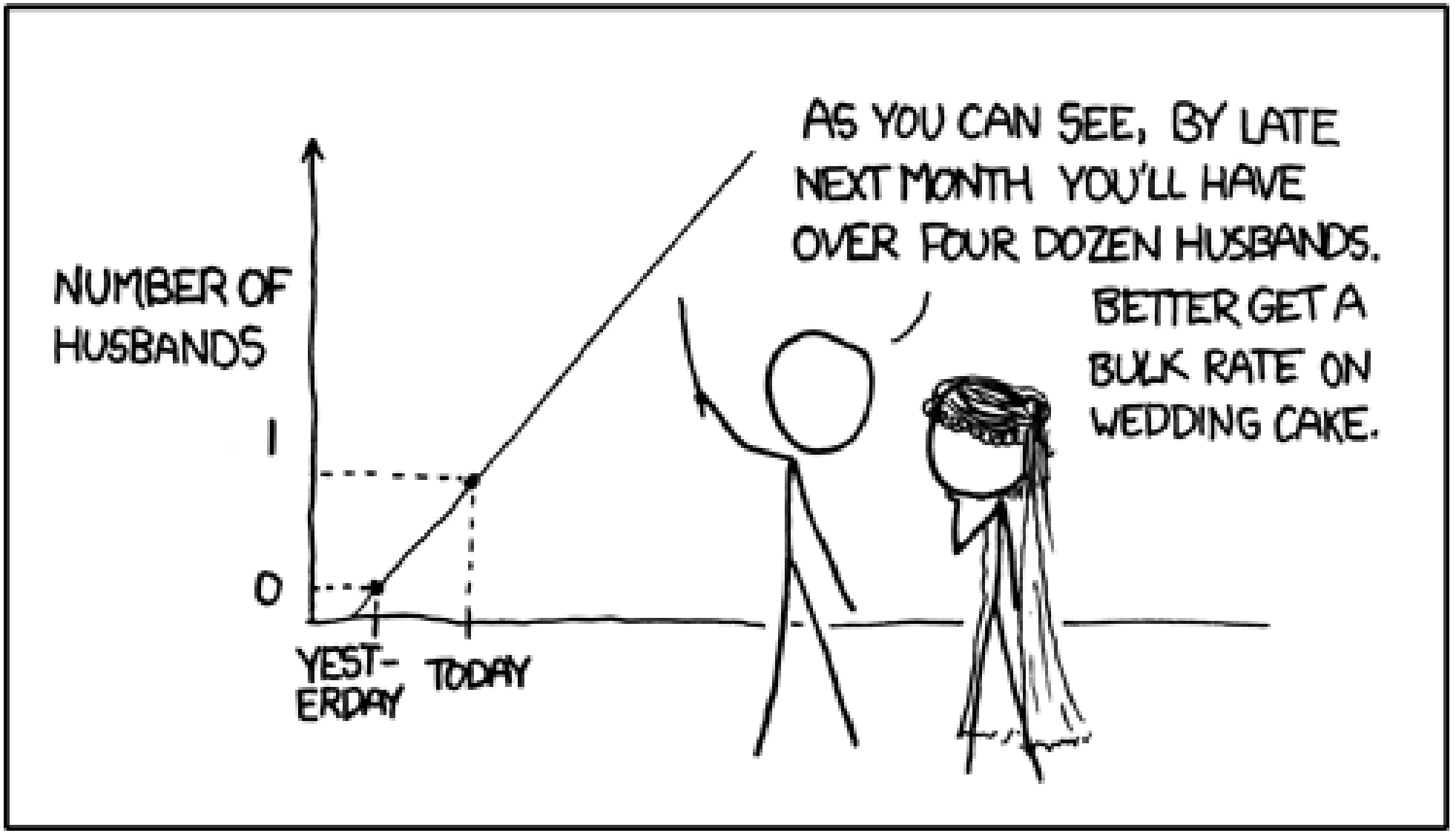
$$BIC = k * ln(N) - 2 * ln(L)$$

  - Akaike Information Criterion:

$$AIC = 2 * k - 2 * ln(L)$$

- The BIC penalizes more the number of parameters than the AIC

MY HOBBY: EXTRAPOLATING



Summary

- Social Impact Theory
  - How a social situation determines the extent of behavior change
  - Hypothesizes a multiplicative effect of the number of sources, their strength, and their immediacy
  - Extends to one source and an audience with division of impact

- Social Influence in Online Media
  - The origin of the viral analogy and its limitations
  - Examples of Digg and Reddit and where SIT starts to fail

- Linear Regression in SMDA
  - How to estimate a quantity as a linear combination of others
  - Goodness of fit: R squared and likelihood-based metrics