Information Theoretic Approaches for Understanding Human Behavior

Greg Ver Steeg and Aram Galstyan

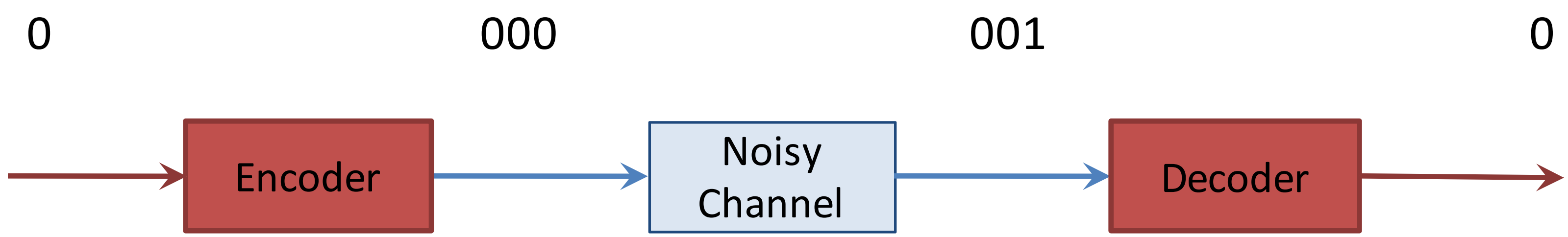
University of Southern California
Information Sciences Institute

March 9, 2015
IPAM Tutorial

USC
VITERBI
SCHOOL OF
ENGINEERING

ISI
Information Sciences Institute

Information theory:
Reliable communication over a noisy channel



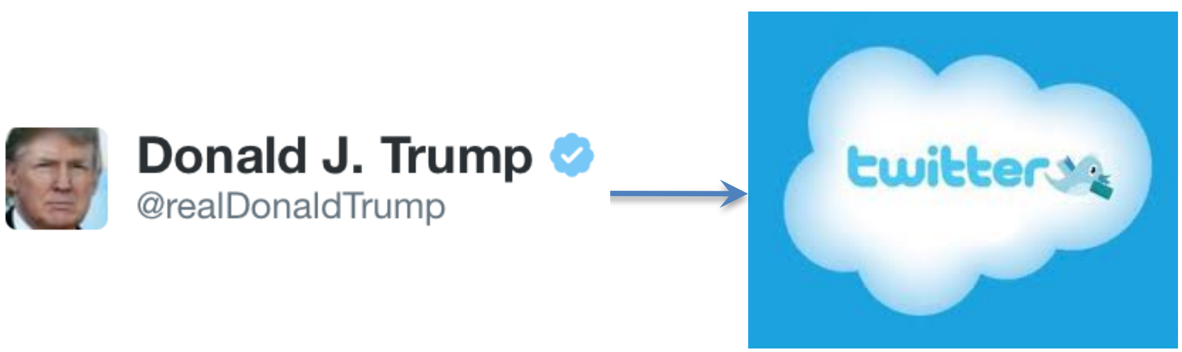
How much information can we send?

What is the maximum rate of *error-free* communication over *all possible codes*?

Surprises:

- Error free is possible!
- Simple formula for this rate! (Mutual information)

Examples of noisy channels



The Bandwagon

CLAUDE E. SHANNON

1956

"Information theory has, in the last few years, become something of a scientific bandwagon…

It will be all too easy for our somewhat artificial prosperity to collapse overnight when it is realized that the use of a few exciting words like *information*, *entropy*, *redundancy* do not solve all of our problems"

In the context of “culture analytics”,

our problems are:

- *Useful, meaningful measures*
- *Estimation*

- Information Theory Basics
  - Entropy, MI, Discrete IT estimators
  - Entropy estimation demo
- Human behavior dynamics
  - Social networks
  - Stylistic coordination

Coffee Break (3:15-3:30)

- Non-parametric entropy estimation
- Very high-dimensional information
  - How to handle it?
  - Applications: language, personality, behavior

Basics

- Plain Old Entropy
  - Why "log"?, Building intuition
  - Continuous variable caveats
- Mutual information
  - Definition/interpretation/forms
  - Continuous variables
  - Dependence/multivariate measures
- Estimation for discrete variables

Why "log"?

- A random variable $$p(X = x) = p(x) = 1/6$$ $$x = 1, \ldots, 6$$

- How would we quantify uncertainty, H(X)?
- 2 dice: 6*6 = 36 states
- log(6*6) = log(6) + log(6) = 2 log(6)

Axiomatic approach (Shannon)

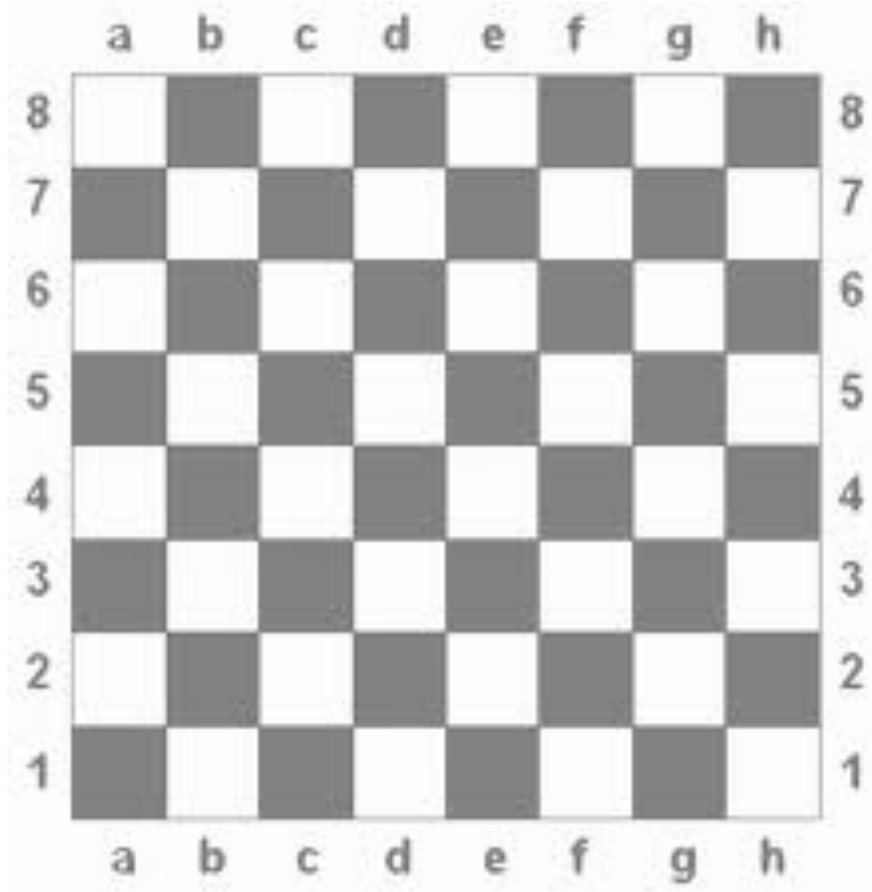
- Which functions quantify uncertainty?
  - **Continuous** (a small change in p(x) should lead to a small change in our uncertainty)
  - **Increasing** (If there are n equally likely outcomes, uncertainty goes up with n)
  - **Composition** (The uncertainty for two independent coins should equal the sum of uncertainties for each coin)

$$
\begin{aligned}
H(X) &= \mathbb{E}(\log 1/p(x)) \\
&= -\sum_x p(x) \log p(x)
\end{aligned}
$$

Alternate interpretation: compression

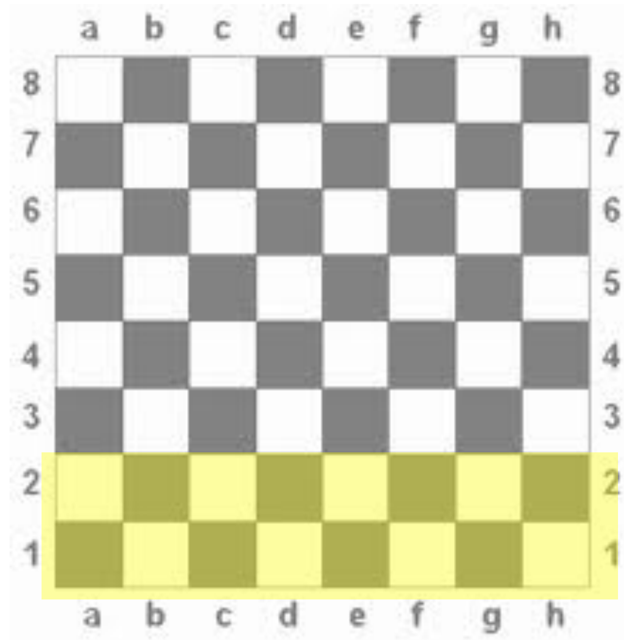
Guess my square game:

- I pick a square uniformly at random
- You can ask yes/no questions to determine the square

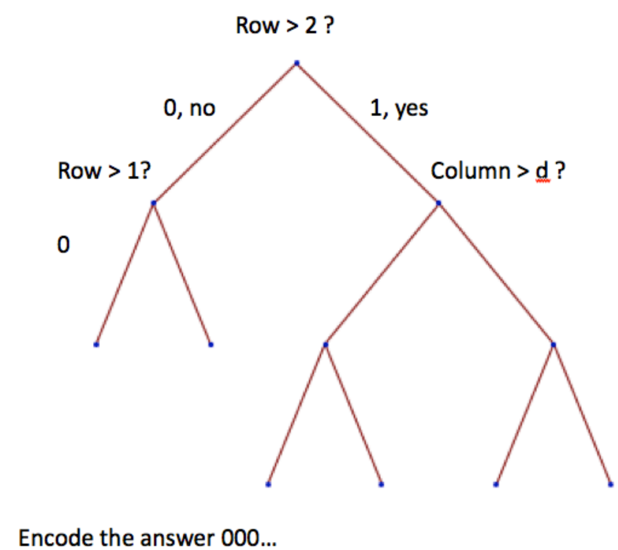


- How many questions are required?
- To distinguish between **N** squares, we need **$\log_2$ N** questions

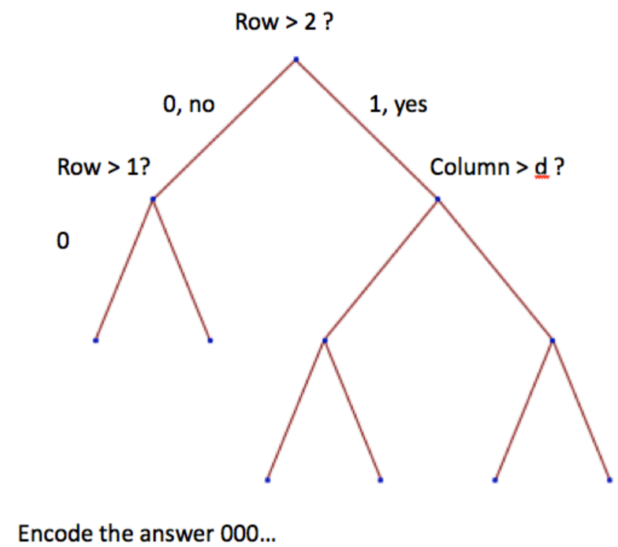
- In Round 2: I prefer the bottom two rows, and half the time pick one of those squares



- Find the correct square with fewer questions *on average*

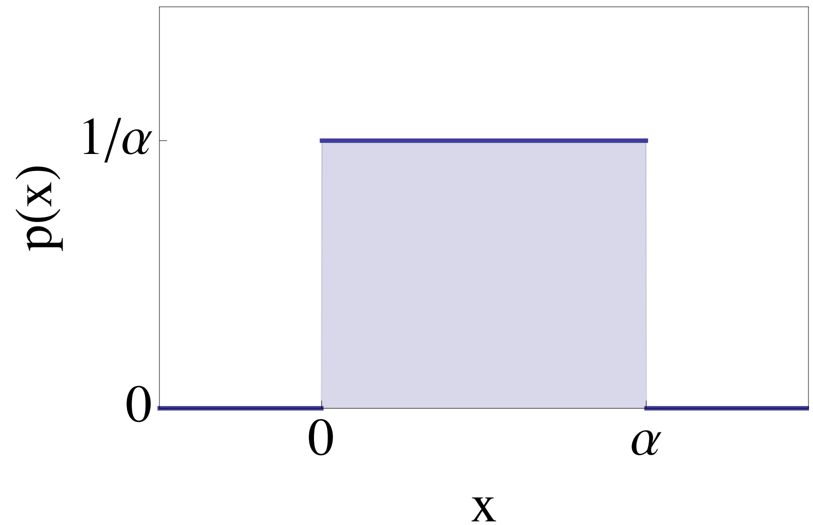


- How many questions do we need *on average?*
- This answer is exactly the entropy and therefore entropy can be viewed as a measure of *compression*



Continuous Random Variables (are a little different)

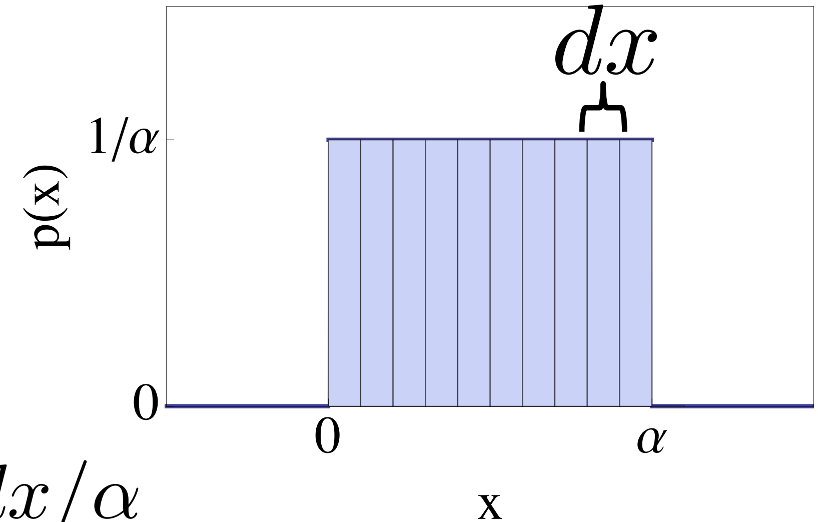
- A probability density



- What is the probability of observing x=0.53243289750432856390573 2…?
- p(x)dx tells us the probability observe a number in [x,x+dx)

(Differential) Entropy

- p(x)dx tells us the probability observe a number in [x,x+dx)



Each discrete bin has probability dx/α

$$H(X) = -\sum_{i=1}^{\alpha/dx} dx/\alpha \log dx/\alpha$$

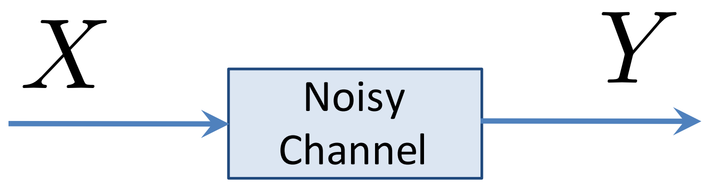
$$= \log\alpha - \cancel{\log dx}$$

As $dx \to 0 \ldots$

$$H_{diff}(X) = \int dx\ p(x) \log p(x) = \mathbb{E}(\log 1/p(x))$$

- Plain Old Entropy
  - Why “log”?, Building intuition
  - Continuous variable caveats
- **Mutual information**
  - Definition/interpretation/forms
  - Continuous variables
  - Dependence/multivariate measures
- Estimation for discrete variables

Mutual information



$$C = \max_{p(X)} I(X : Y)$$

Mutual information!

Channel Coding Theorem (Shannon, 1948)

For every $R < C$, there are channel codes that allow almost error-free transmission of information.

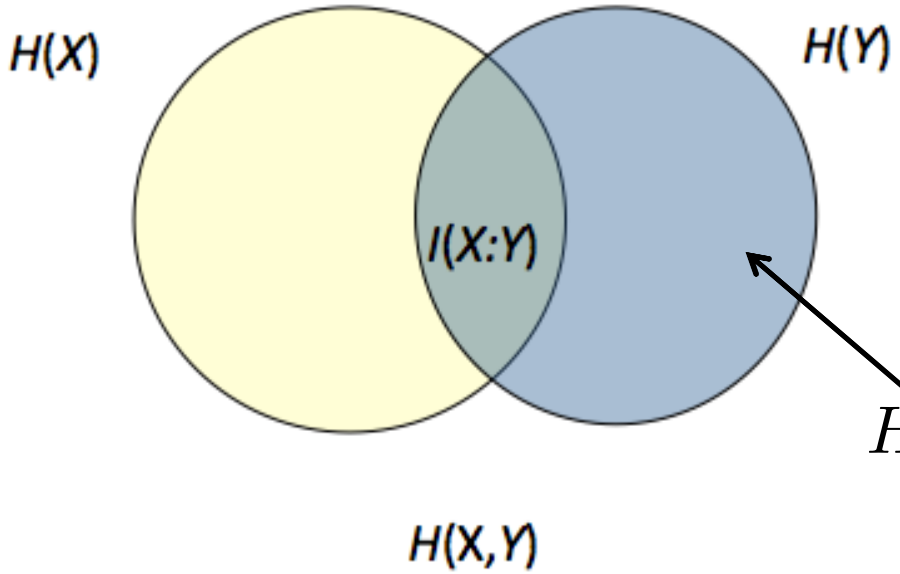
Mutual information

$$I(X:Y) = \underbrace{H(X) + H(Y)}_{\text{Uncertainty if X and Y are independent}} - \underbrace{H(X,Y)}_{\text{Uncertainty considered as one system}}$$

Some things to notice:

- Symmetric
- A difference of entropies
- Non-negative

Mutual information



$$H(Y|X) = \sum_x p(x) H(Y|X=x)$$

Read off other the ways of describing mutual information:

$$\begin{aligned} I(X:Y) &= H(X) + H(Y) - H(X,Y) \\ &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \end{aligned}$$

Independence

$$I(X:Y) = \underbrace{H(X) + H(Y)}_{\text{Uncertainty if X and Y are independent}} - \underbrace{H(X,Y)}_{\text{Uncertainty considered as one system}}$$

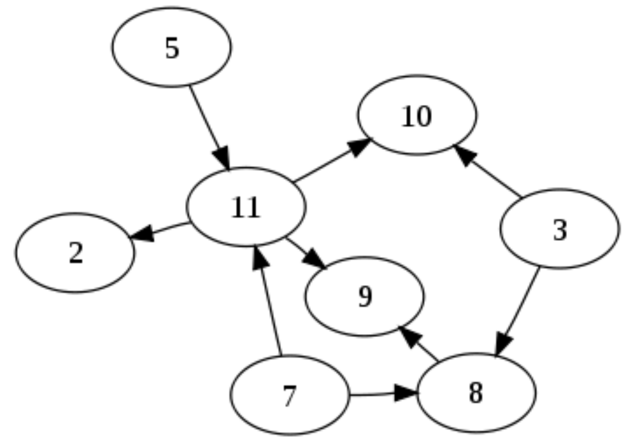
$$H(X) = \mathbb{E}\left(\log 1/p(x)\right)$$

$$\begin{aligned} I(X:Y) &= \mathbb{E}\left(\log 1/p(x) + \log 1/p(y) - \log 1/p(x,y)\right) \\ &= \mathbb{E}\left(\log \frac{p(x,y)}{p(x)p(y)}\right) \end{aligned}$$

$$I(X:Y) = 0 \longleftrightarrow p(x,y) = p(x)p(y)$$

Extends to Conditional Independence

- Bayesian networks, e.g., can be read as encoding a set of "conditional independence" relationships



$$p(X,Y|Z) = p(X|Z)p(Y|Z)\forall Z \longleftrightarrow X \perp Y|Z$$

$$X \perp Y|Z \longleftrightarrow I(X:Y|Z) = 0$$

$$I(X:Y|Z) = H(X|Z) - H(X|Z,Y)$$

First useful(?) property for M.L.

$$I(X:Y)=0 \longleftrightarrow p(x,y)=p(x)p(y)$$

- You don't get this for other "correlation" measures: (Pearson, Kendall, Spearman…)
- MI captures nonlinear relationships, the size of MI has many nice interpretations
- Extends to multivariate (**last part**)
- But, is it "useful"? It depends on p(x,y)…

- Plain Old Entropy
  - Why “log”?, Building intuition
  - Continuous variable caveats
- Mutual information
  - Definition/interpretation/forms
  - Continuous variables
  - Dependence/multivariate measures
- **Estimation for discrete variables**

Estimation for discrete variables

- An "asymptotically unbiased" estimator:

$$x^{(i)} \sim p(X), i = 1, \ldots, N$$

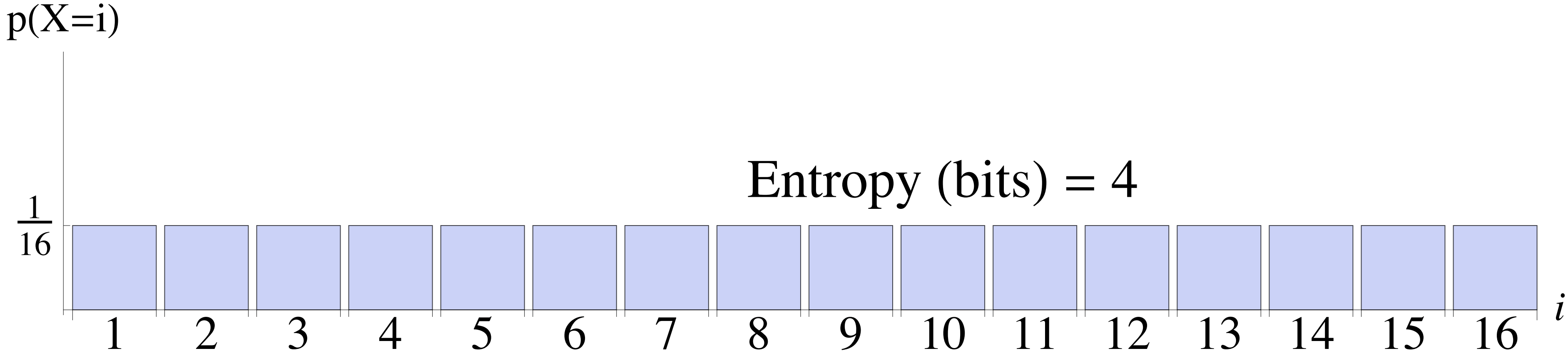
$$\lim_{N \to \infty} \mathbb{E}\left[\hat{H}_N(X)\right] = H(X)$$

- For discrete entropy, the 'plug-in' estimator:

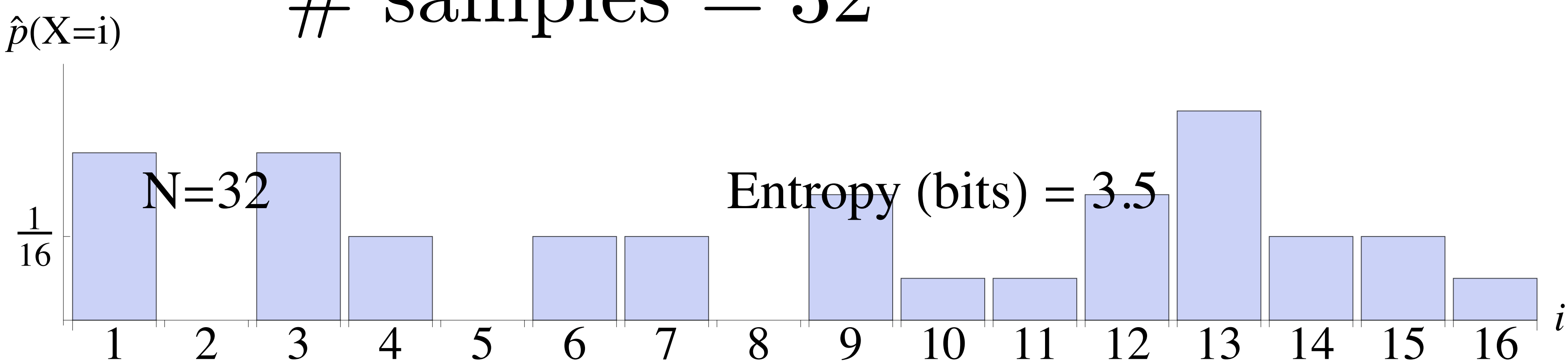
$$\hat{H}(X) = -\sum_x \hat{p}(x) \log \hat{p}(x)$$

$$\hat{p}(x) = (\text{number of times to observe } x)/N$$

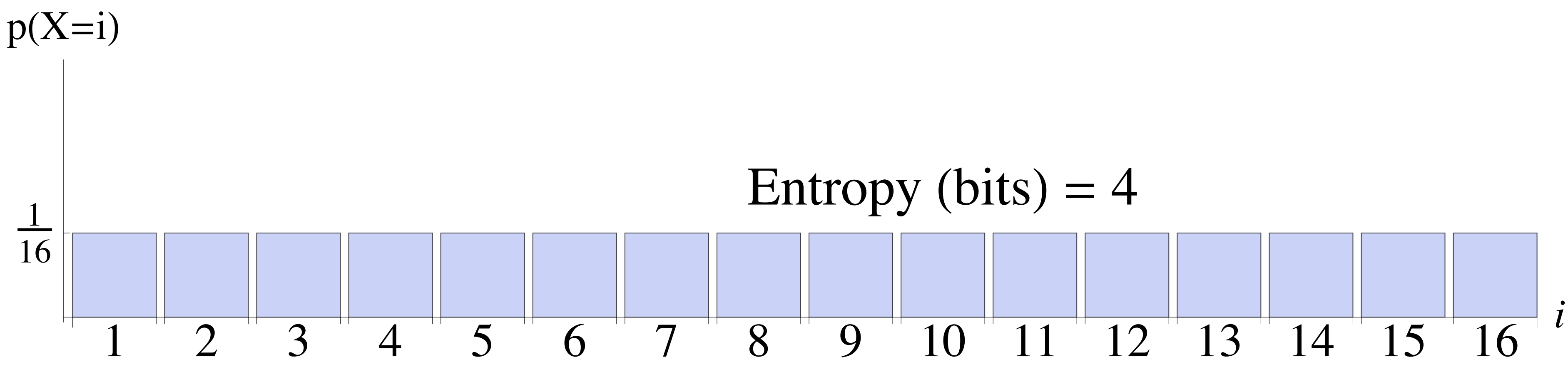
How well do we do?

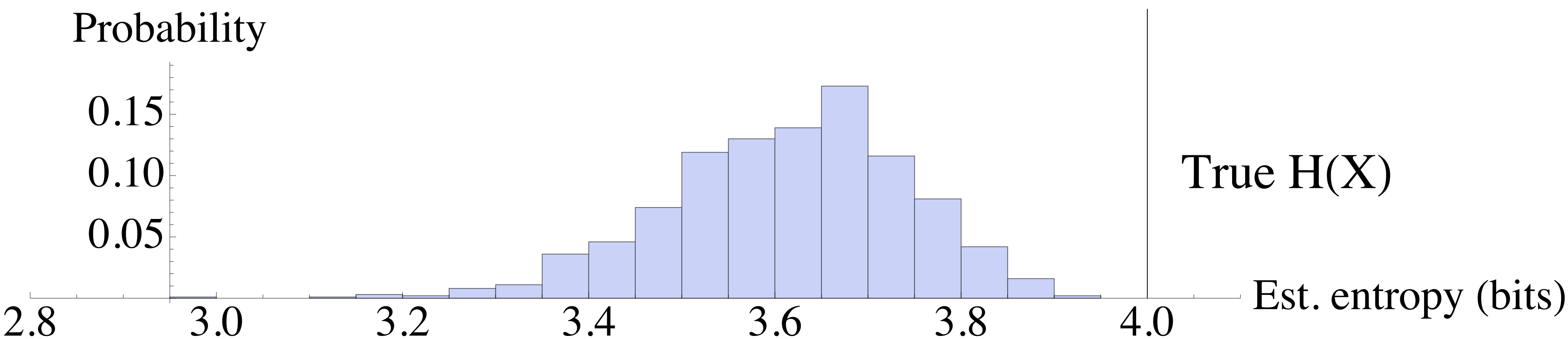


$$\begin{aligned} \#\ \text{states} &= 16 \\ \#\ \text{samples} &= 32 \end{aligned}$$



How well do we do?





$\#$ states $= 16$
$\#$ samples $= 32$

Naïve estimator for MI?

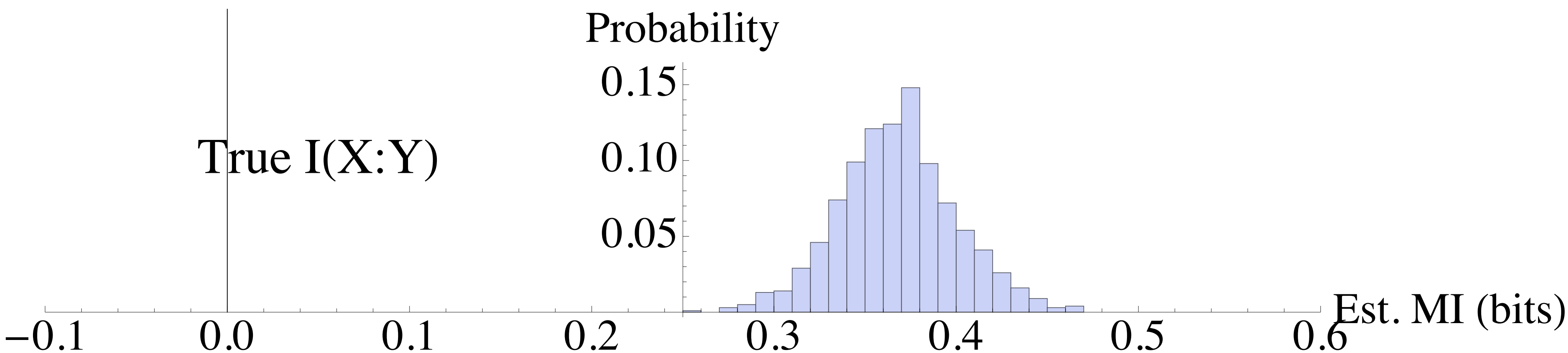
Again, standard formula using observed freq. counts:

$$\hat{I}(X:Y) = \mathbb{E}\left(\log \frac{\hat{p}(x,y)}{\hat{p}(x)\hat{p}(y)}\right)$$

Bias for MI

E.g., for $x = 1, \ldots, 16$ and $y = 1, \ldots, 16$
$p(x, y) = 1/(16 \cdot 16)$
Then $I(X : Y) = 0$.

Again, let $\#$ samples $= 2\cdot$ $\#$ states

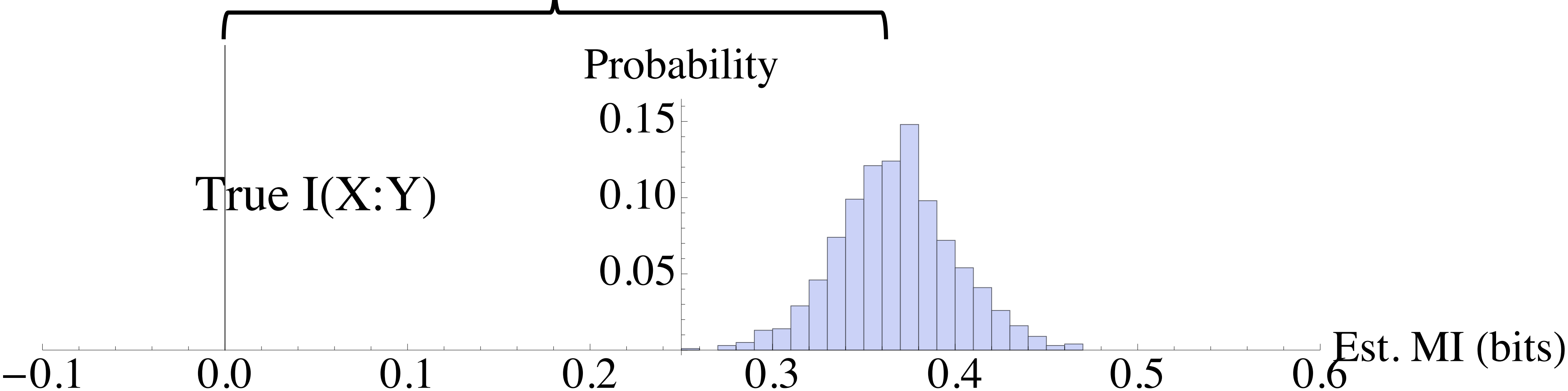


Three possible solutions

- Analytic estimate of bias (Panzeri-Treves)
- Bootstrap
- Shuffle Test

Bias for MI

$$\propto \frac{\#\text{states}}{\#\text{samples}}$$



Correcting for the Sampling Bias Problem in Spike Train Information Measures

Stefano Panzeri, Riccardo Senatore, Marcelo A. Montemurro and Rasmus S. Petersen

J Neurophysiol 98:1064-1072, 2007. First published 5 July 2007; doi:10.1152/jn.00559.2007

Bias for MI

- Bootstrap: generate new samples based on $\hat{p}(x, y)$
- Estimate bias for those samples, use as correction

Entropy **2013**, *15*, 2246-2276; doi:10.3390/e15062246

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

Bootstrap Methods for the Empirical Study of Decision-Making and Information Flows in Social Systems

Simon DeDeo [1,*], Robert X. D. Hawkins [1,2], Sara Klingenstein [1] and Tim Hitchcock [3]

Permutation test

- For a given set of samples

$$(x^{(i)}, y^{(i)}), i = 1, \ldots, N$$

- Generate many "shuffled" versions

$$(x^{\pi(i)}, y^{(i)}), i = 1, \ldots, N$$

- For these, $I(X_{shuffle}, Y) = 0$ this gives empirical CI for correlations to be due to chance.

- Information Theory Basics
  - Entropy, MI, Discrete IT estimators
  - Entropy estimation demo
- **Human behavior dynamics**
  - Social networks
  - Stylistic coordination

Coffee Break (3:15-3:30)

- Non-parametric entropy estimation
- Very high-dimensional information
  - How to handle it?
  - Applications: language, personality, behavior

Information Theoretic Approaches for Understanding Human Behavior

Greg Ver Steeg and Aram Galstyan

University of Southern California
Information Sciences Institute

March 9, 2015
IPAM Tutorial

USC
VITERBI
SCHOOL OF
ENGINEERING

Information Sciences Institute

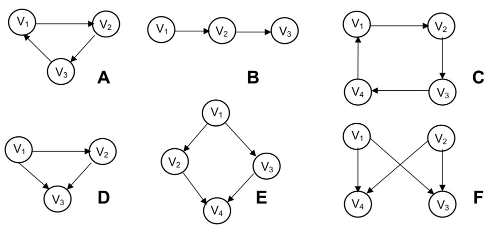
- Information Theory Basics
  - Entropy, MI, Discrete IT estimators
  - Entropy estimation demo
- **Human behavior dynamics**
  - Social networks
  - Stylistic coordination

Coffee Break (3:15-3:30)

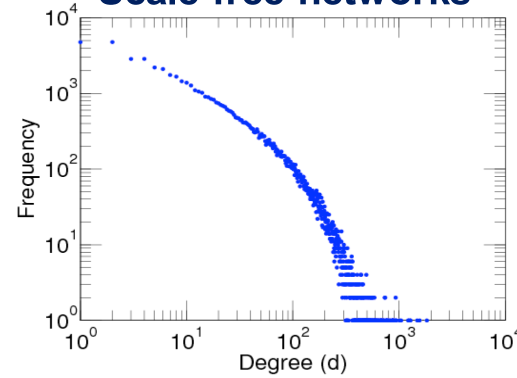
- Non-parametric entropy estimation
- Very high-dimensional information
  - How to handle it?
  - Applications: language, personality, behavior

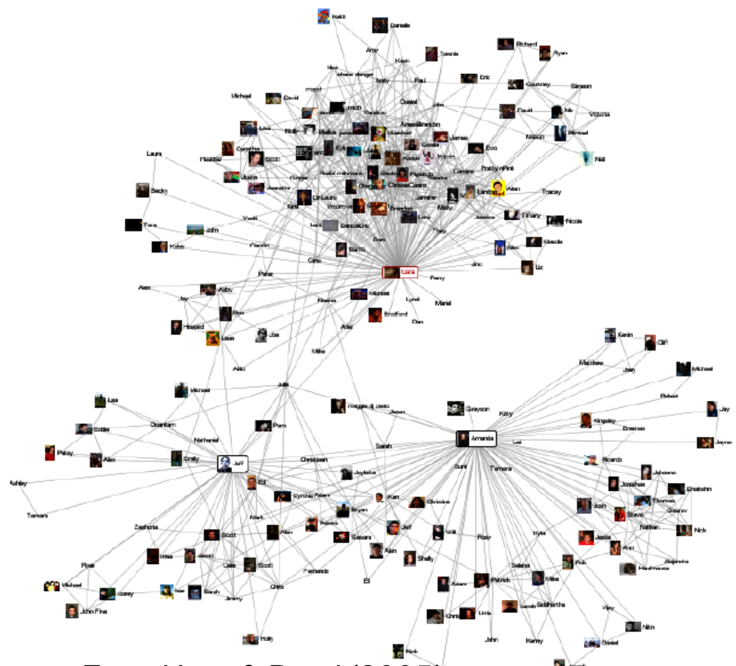
Topology of social interactions

Mesoscopic motifs

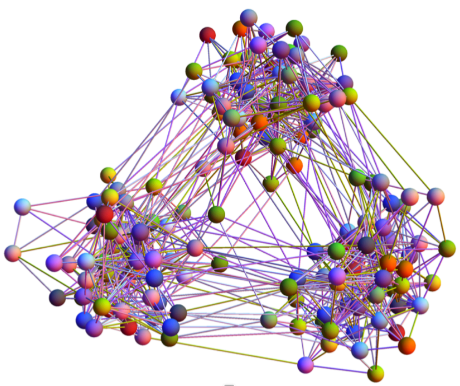


Scale-free networks

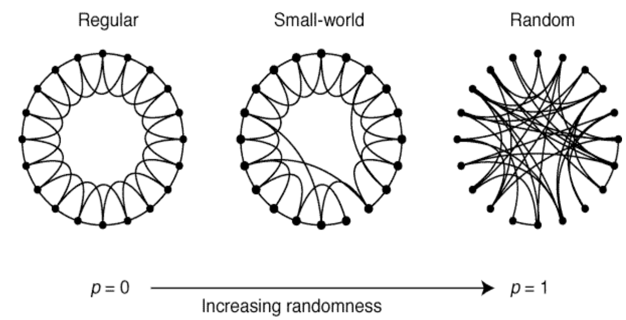




From Heer & Boyd (2005).



Modularity & group structure

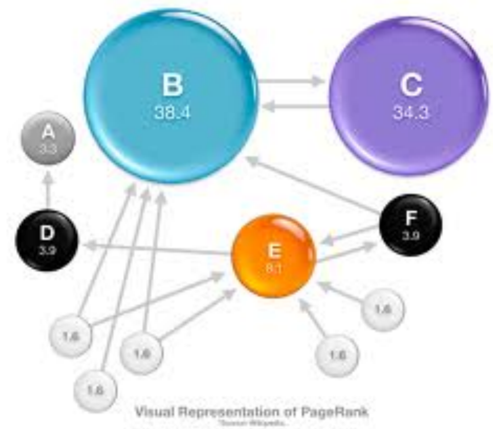


Clustering and Small Worlds

What about behavioral data?

Measuring influence

- Structural (network) measures
  - Out-degree/number of followers
  - Page-rank, other centrality measures
- Does not consider user dynamics
- Not all links are created equal



“Social Capital” on ebay

35,000+ Twitter Followers within 48 Hours!

$14.99
Buy It Now
Free shipping

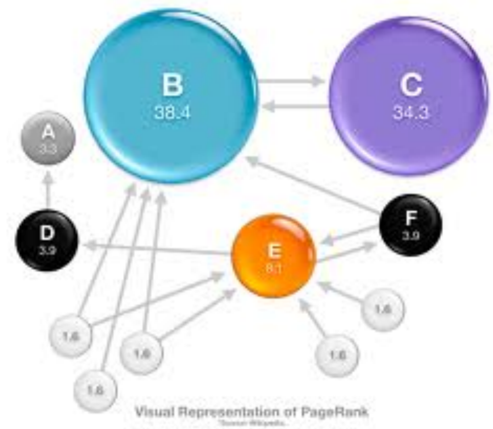
Newly Listed Tweet from Health and Wellness Twitter Handle With 9,000 Followers

2h left
Today 12:46PM

$0.01
1 bid
Free shipping

Measuring influence

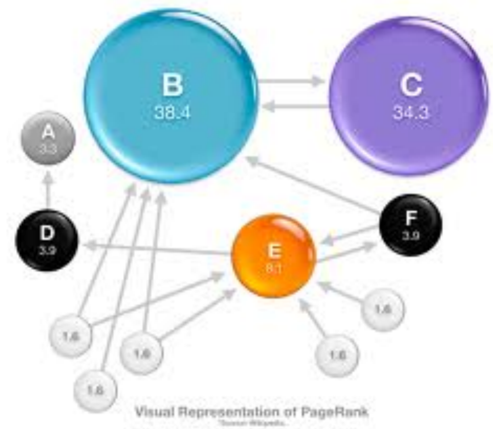
- Structural (network) measures
  - Out-degree/number of followers
  - Page-rank, other centrality measures
- Does not consider user dynamics
- Not all links are created equal



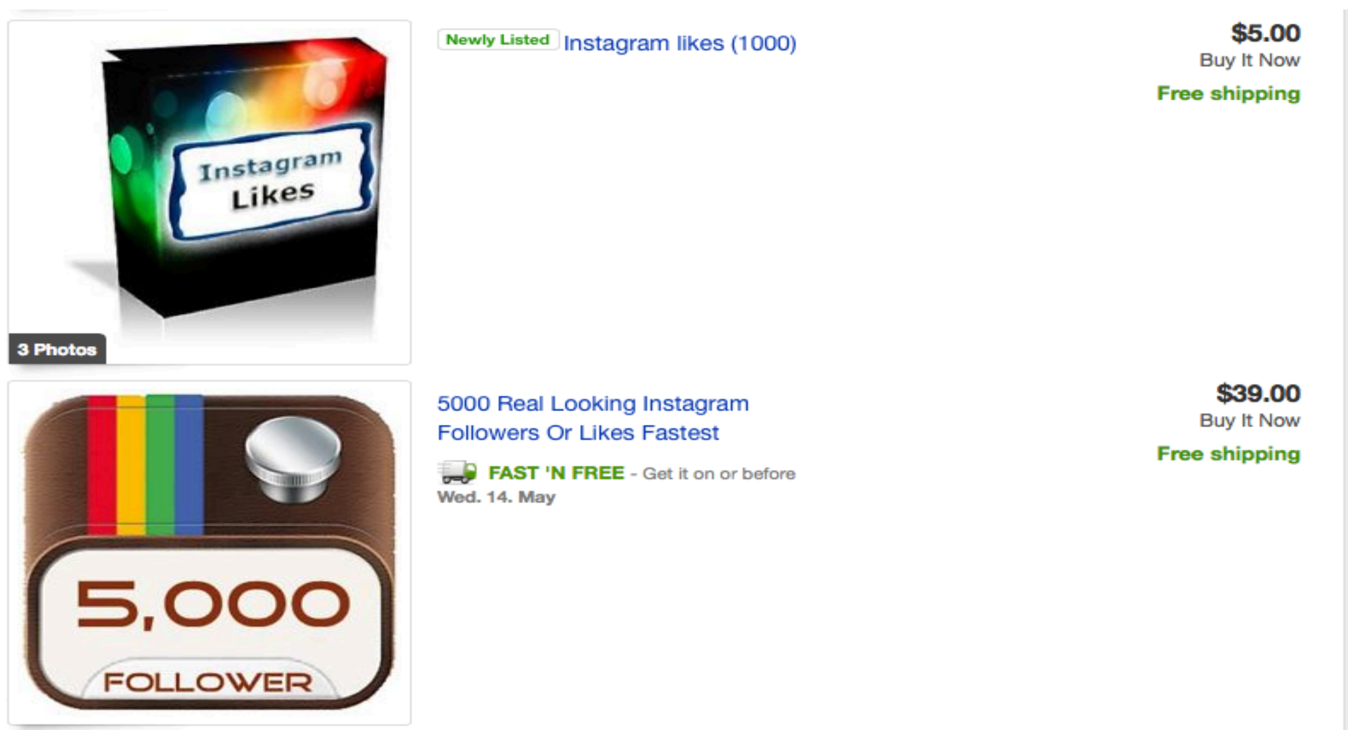
“Social Capital” on ebay

Measuring influence

- Structural (network) measures
  - Out-degree/number of followers
  - Page-rank, other centrality measures
- Does not consider user dynamics
- Not all links are created equal

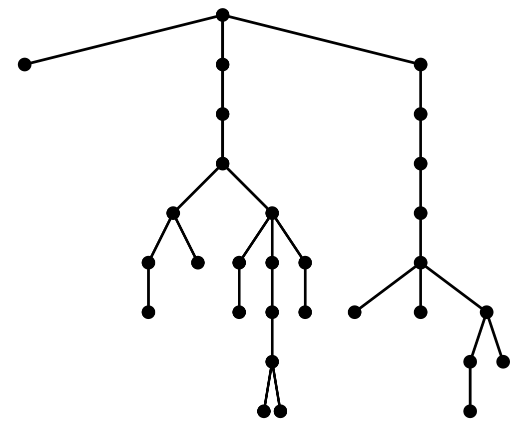


“Social Capital” on ebay



Measuring influence

- Dynamic measures
  - Re-tweets (Kwak et. al. WWW '10)
  - Size of cascades (Bakshy, et. al. WSDM '11)
  - Influence-passivity (Romero et. al. WWW '11)



- Requires explicit causal knowledge
  - E.g, who responds to whom
- Platform-specific
  - *Retweets/mentions/Likes*
- Tailored to particular activity/representation
  - *Text/check-in/purchase/etc*

Influence via predictability

- *Y* influences *X* if Y's past activity is a good predictor of X's future activity

- Quantified using information-theoretic concepts
  - E.g., *Transfer Entropy (Schreiber, 2000)*: How much our uncertainty about user X's future activity is reduced by knowing Y's past activity

$$TE_{Y \to X} = H(X^{\text{Future}} | X^{\text{Past}}) - H(X^{\text{Future}} | Y^{\text{Past}}, X^{\text{Past}})$$

Uncertainty about X

Uncertainty about X, if you know Y's past activity

Model-free

***X, Y* can represent:**
Timing of activity
Location
Content
Style
...

Transfer Entropy

- Entropy of a random variable X

$$H(X) = -\sum_x p(x) \log p(x) \qquad \textbf{discrete}$$

$$-\int dx p(x) \log p(x) \qquad \textbf{continuous}$$

- Mutual Information

$$I(X:Y) = H(X) - H(X \mid Y)$$

- Conditional Mutual Information

$$CMI(X:Y \mid Z) = H(X \mid Z) - H(X \mid Z,Y)$$

$$TE_{Y \to X} = CMI(X^{Future} : Y^{Past} \mid X^{Past})$$

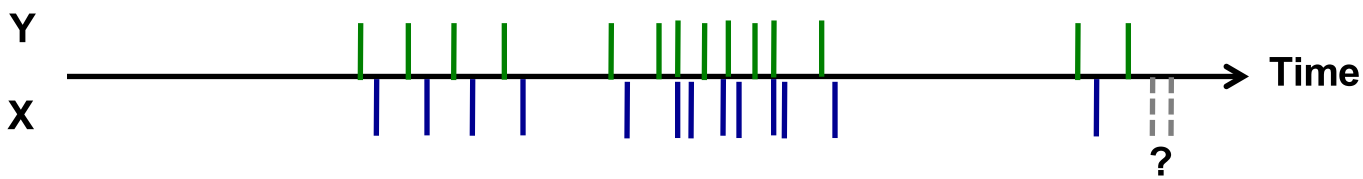
Outline

- Social influence via transfer entropy
  - Activity timing
  - Content dynamics
- Stylistic influence in dialogues
- Estimation of entropic measures (from limited data)

Transfer entropy with activity timing

How predictable is X's behavior? Look at X's history

And if we add Y's history?

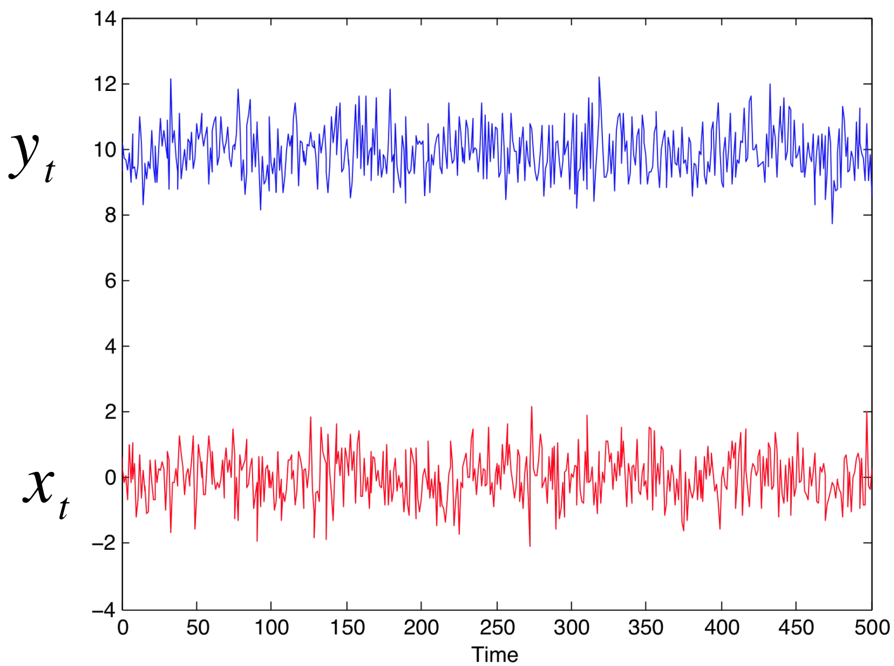


$$TE_{Y \to X} = H(X^{\text{Future}}|X^{\text{Past}}) - H(X^{\text{Future}}|Y^{\text{Past}}, X^{\text{Past}})$$

Uncertainty about X

Uncertainty about X, if you know Y's behavior

Granger Causality



Clive Granger

Model-1 $$x_{t+1} \approx \sum_{j=1}^{p} A_j x_{t-j}$$

Model-2 $$x_{t+1} \approx \sum_{j=1}^{p} A_j x_{t-j} + \sum_{j=1}^{l} B_j y_{t-j}$$

Y is Granger-causal to X if Model-2 is better than Model-1

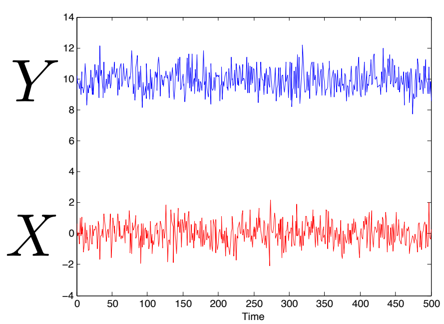
Uncovering Networks from Activities

- **Information transfer** (Schrieber, 2000)
  - How much is our uncertainty about user X's future activity reduced by knowing about Y's past activity?

$$IT_{Y \to X} = \underbrace{H(X^{Future} | X^{Past})}_{\text{Uncertainty about X}} - \underbrace{H(X^{Future} | X^{Past}, Y^{Past})}_{\text{Uncertainty about X, if you know Y}}$$

    - *Arbitrary signals/relationships; hard to evaluate*

- **Granger Causality** (Granger, 1969)



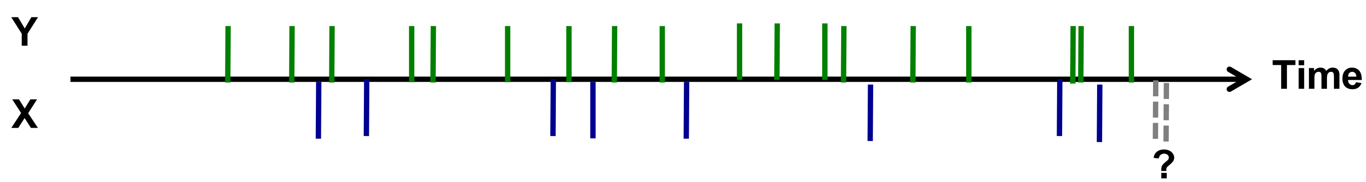
$$(M1) \quad X_{t+1} = \sum_{j=1}^{p} A_j X_{t-p}$$

$$(M2) \quad X_{t+1} = \sum_{j=1}^{p} A_j X_{t-p} + \sum_{k=1}^{m} B_j Y_{t-k}$$

  - **Y** is "Granger-causal" to **X** if (M2) is a better predictor than (M1)
    - *More efficient but assumes linearity; real-valued signals only*

More intuition about T.E.

Alternate possibility: low transfer entropy



Both Deterministic

Both Random

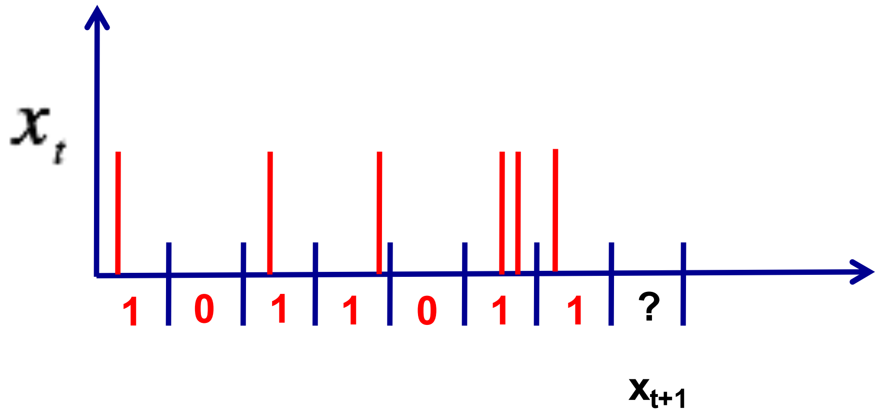
$$TE_{Y \to X} = H(X^{\text{Future}} | X^{\text{Past}}) - H(X^{\text{Future}} | Y^{\text{Past}}, X^{\text{Past}})$$

Uncertainty about X

Uncertainty about X, if you know Y's behavior

Information theory of spike trains

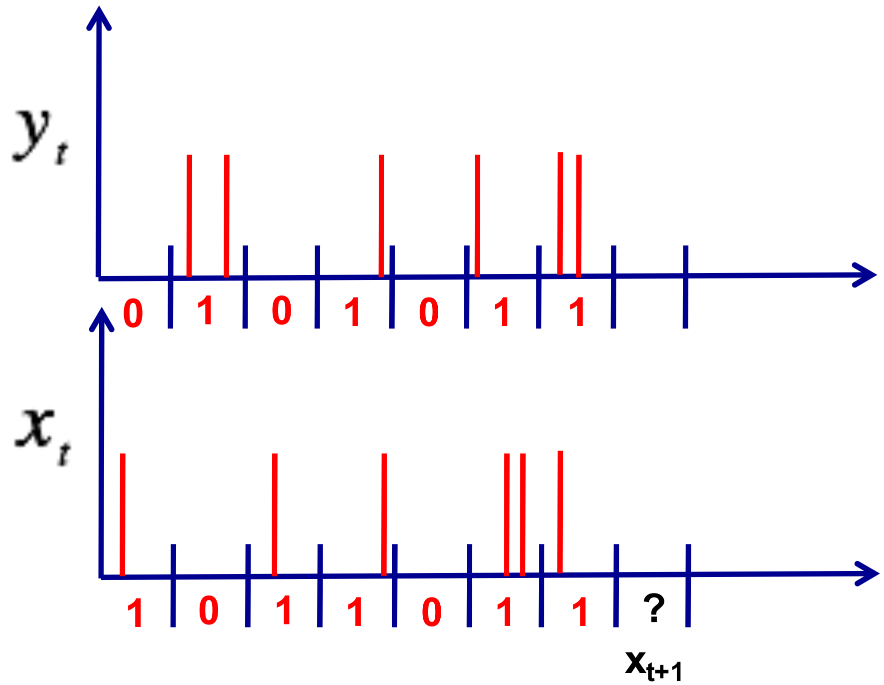
• Information theory has been used for decoding electrical signals in the brain, called "spike trains"



$$p(x_{t+1} \mid x_t^{(k)}) \qquad x_t^{(k)} = x_t, x_{t-1}, \ldots, x_{t-k}$$

$$H(x_{t+1} \mid x_t^{(k)}) = -\sum p(x_{t+1}, x_t^{(k)}) \log p(x_{t,} x_t^{(k)}) / p(x_t^{(k)})$$

How do we calculate this?



$$IT_{Y \to X} = H(x_{t+1} \mid x_t^{(k)}) - H(x_{t+1} \mid y_t^{(k)} x_t^{(k)})$$

1 bit of information transfer from y means we can use y to perfectly predict the next bit of x

Sampling problems

k bins ➔ 2^k possible histories, requiring $O(2^k)$ data

Too little data leads to systematic bias in entropy estimates (Panzeri, et. al. J. Neurophys. 2007)

✓ Get more data/remove inactive users

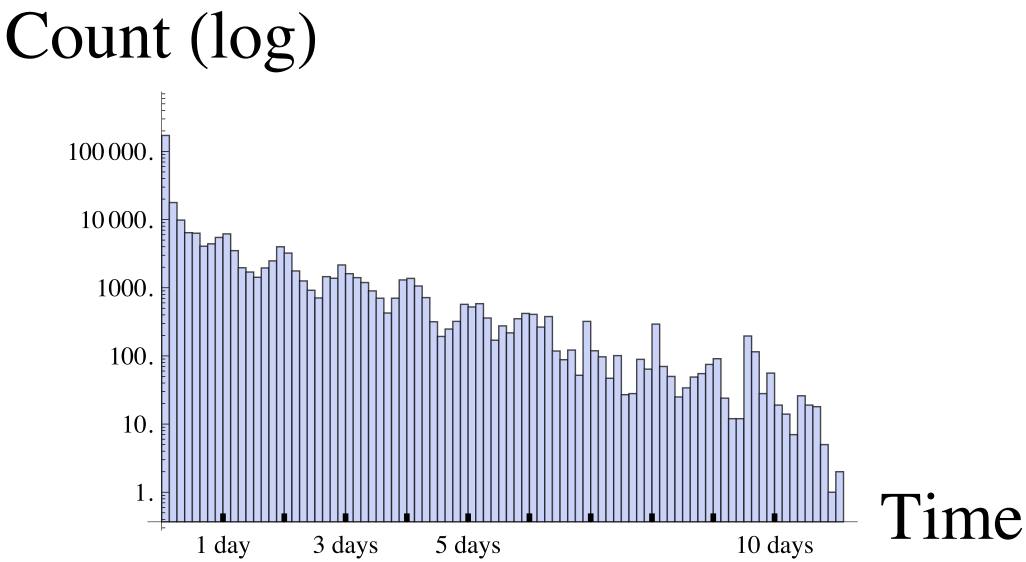
✓ Estimate bias and correct (Panzeri & Treves, 1996)

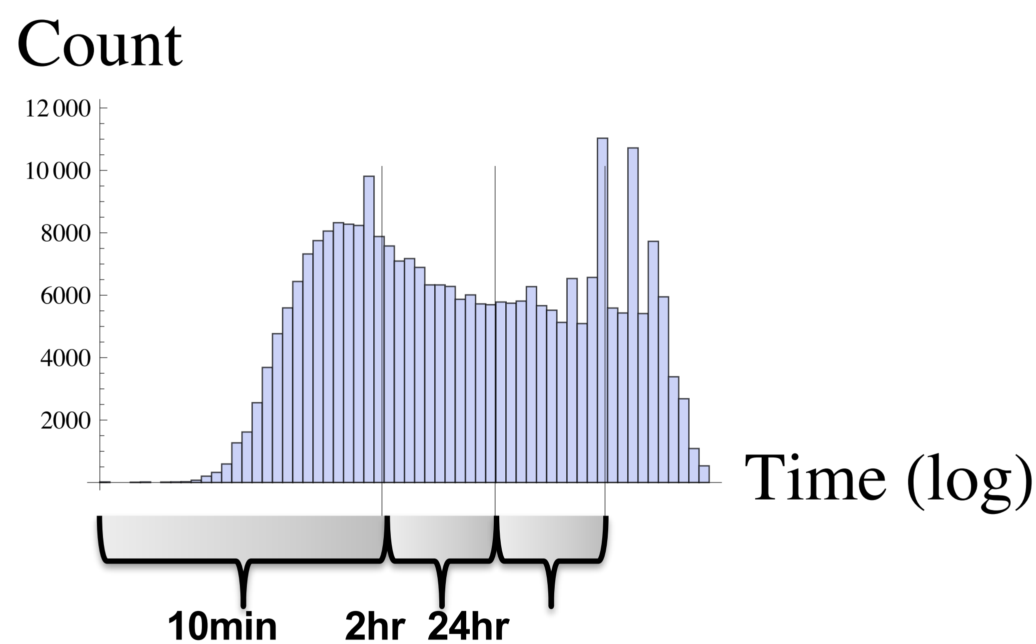
- Use binless, unbiased entropy estimators (Victor, 2003)

✓ Use fewer, more informative bins (for social media)

Relevant time-scales for social media

Histogram of Time to Re-tweet





Results

- Synthetic data
  - *How well can we estimate IT?*
  - *Recover network structure from activity pattern*

- Twitter data
  - *Compare IT to other measures of aggregate influence*
  - *Identify most predictive edges*
  - *IT among top users*
  - *Fine-grain picture of influence*

Synthetic data

Model user activity for two friends, x,y, as a non-homogeneous Poisson process

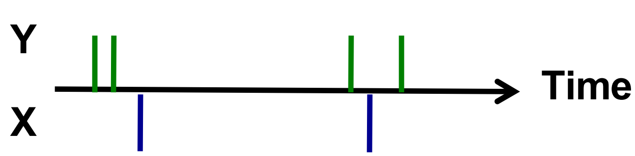
$$\lambda_x(t) = \mu + \gamma \sum_{t_i^y < t} g(t - t_i^y)$$

Rate of activity for user x.

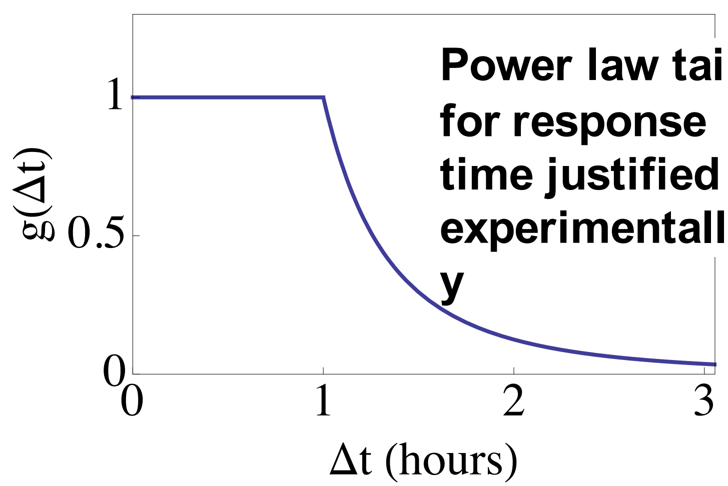
Background rate 1 post/day

Influence strength

Dependence on y's recent posts

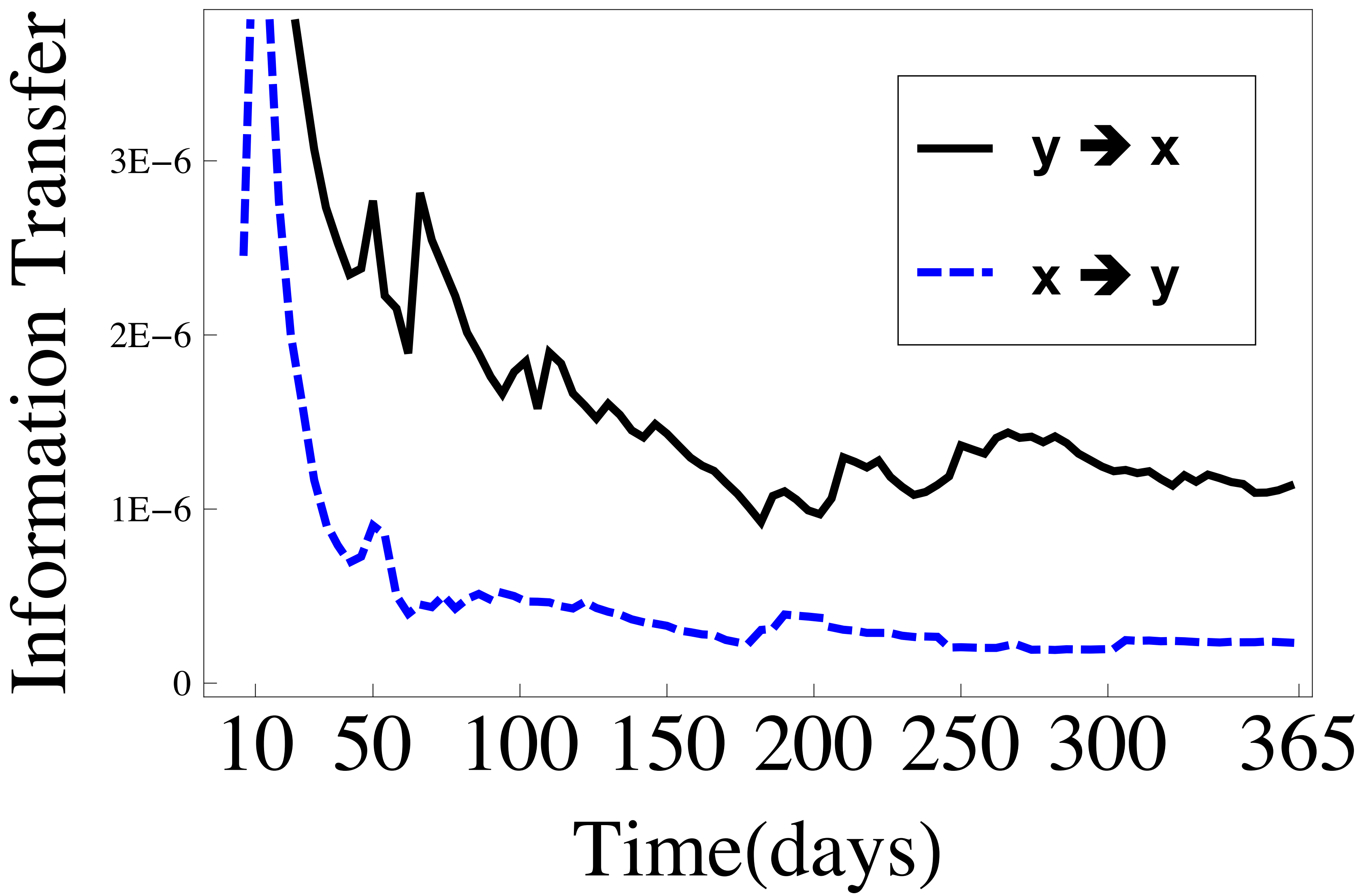


Total time observed, T



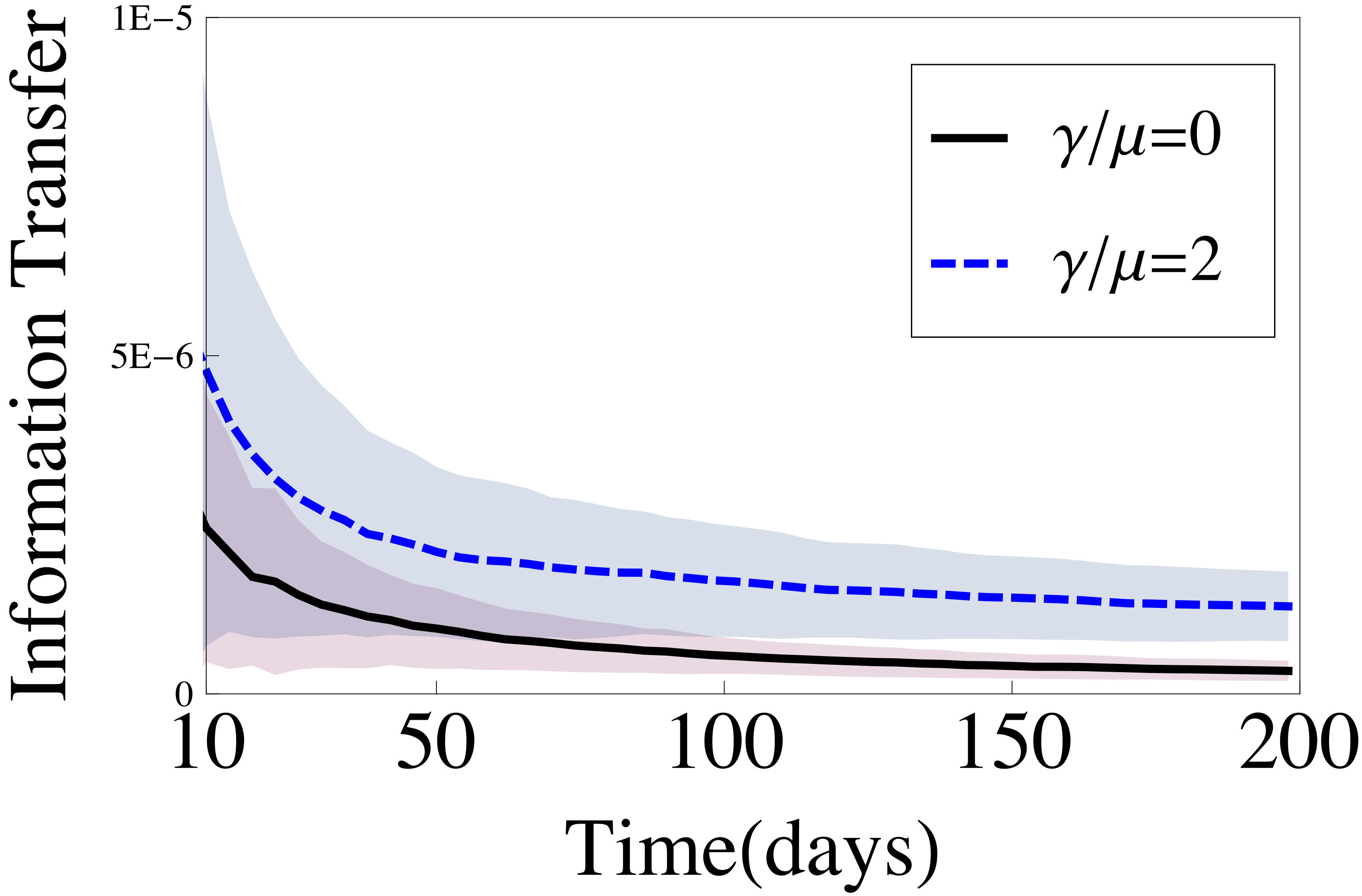
Synthetic data

- If X is affected by Y, but not vice versa, this asymmetry is captured using information transfer



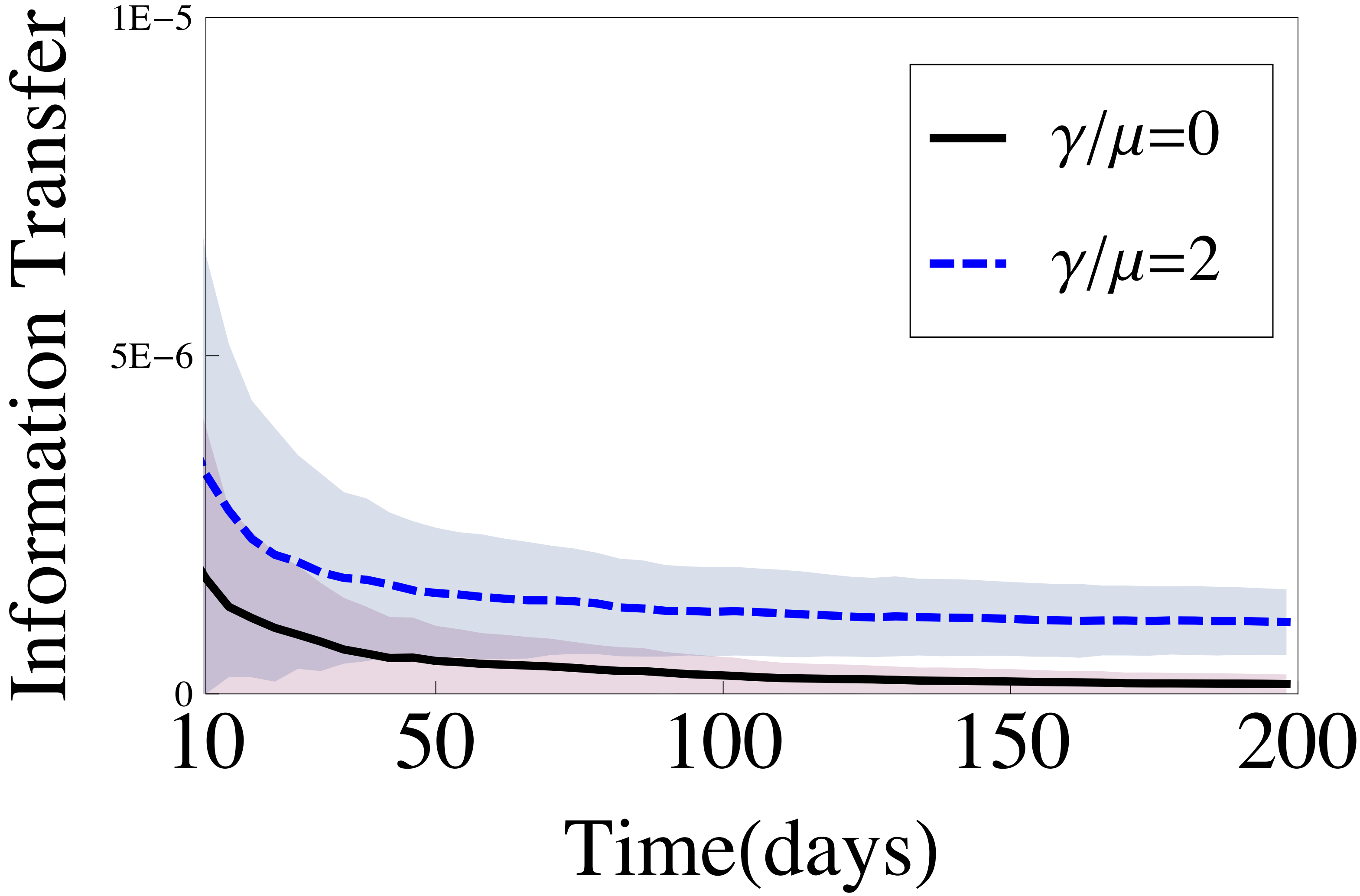
Synthetic data

- Information transfer as a function of how long we observe
- Equivalently, fix time and change the rate of activity

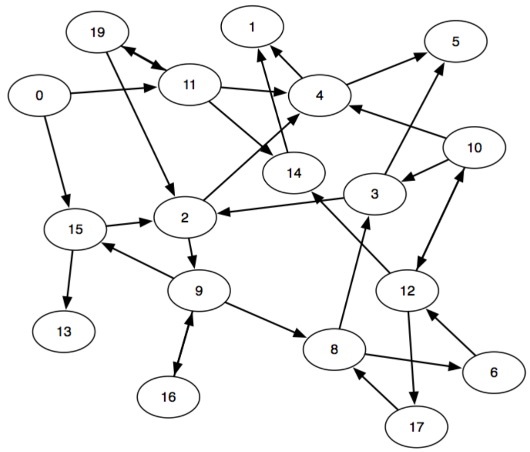


Synthetic data

- Post-bias correction

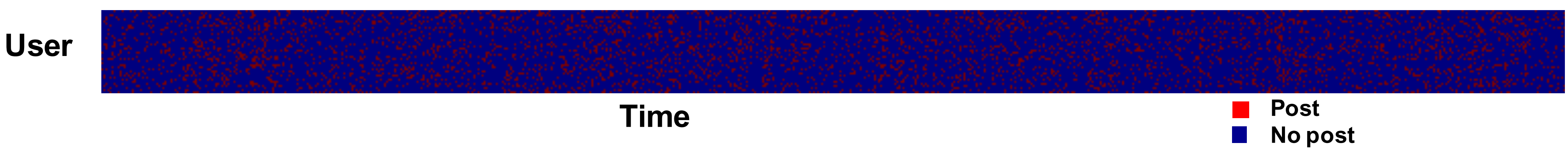


Synthetic data



+

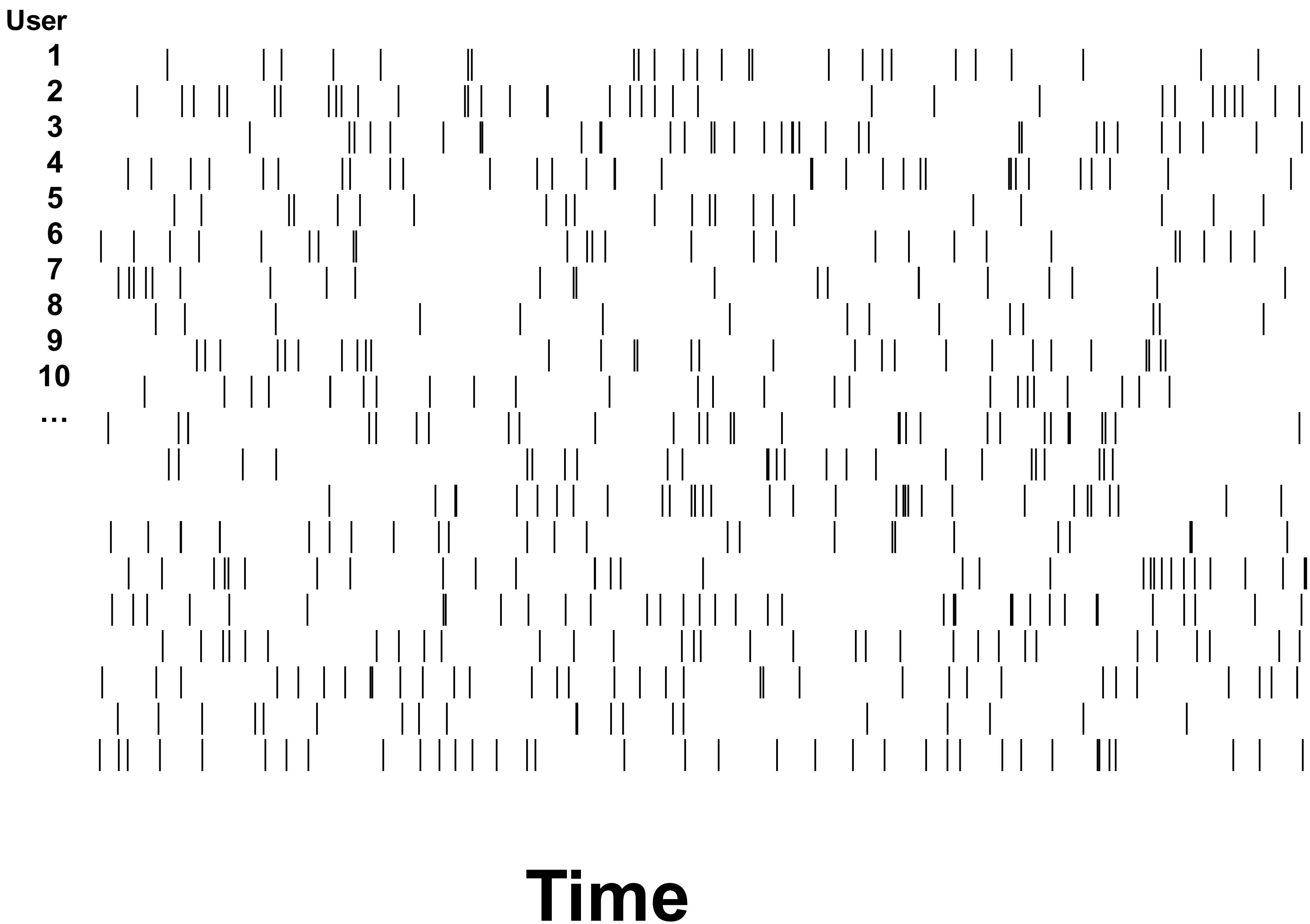
Generate activity according to graph (30 days, background rate = 1 post/day, γ=μ)



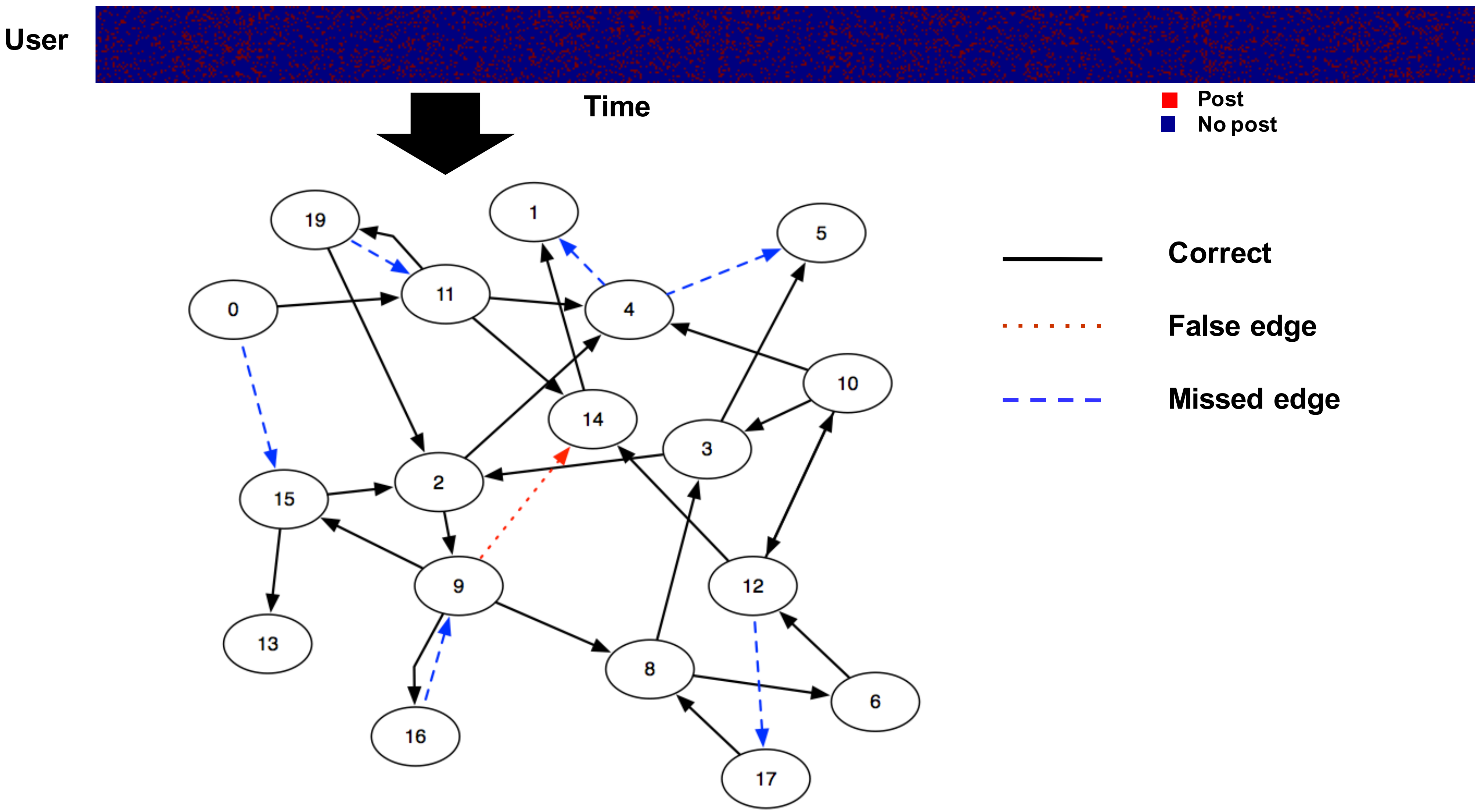
Calculate information transfer between each pair of users.

Can we use this information to recover the correct network?

Who influences whom?



Synthetic data



~ 50 posts/person typically leads to perfect reconstruction of network.

Twitter data

- Top information transfer edges

		Banned	
Free2BurnMusic	➔	~~Free2Burn~~	0.00433
Earn_**Cash**_Today	➔	~~income_ideas~~	0.00116
BuzTweet_com	➔	scate	0.00100
Kamagra_**drug**2	➔	sogradrug3	0.000929
Sougolinkjp	➔	sogolinksite	0.000907
kcal_**bot**	➔	FF_kcal_bot	0.000903
Nr1topforex	➔	~~nr1forex**money**~~	0.000797
Wpthemeworld	➔	wpthememarket	0.000711
Viagrakusurida	➔	viagrakusuride	0.000680
BoogieFonzareli	➔	~~Nyce_Hunnies~~	0.000677

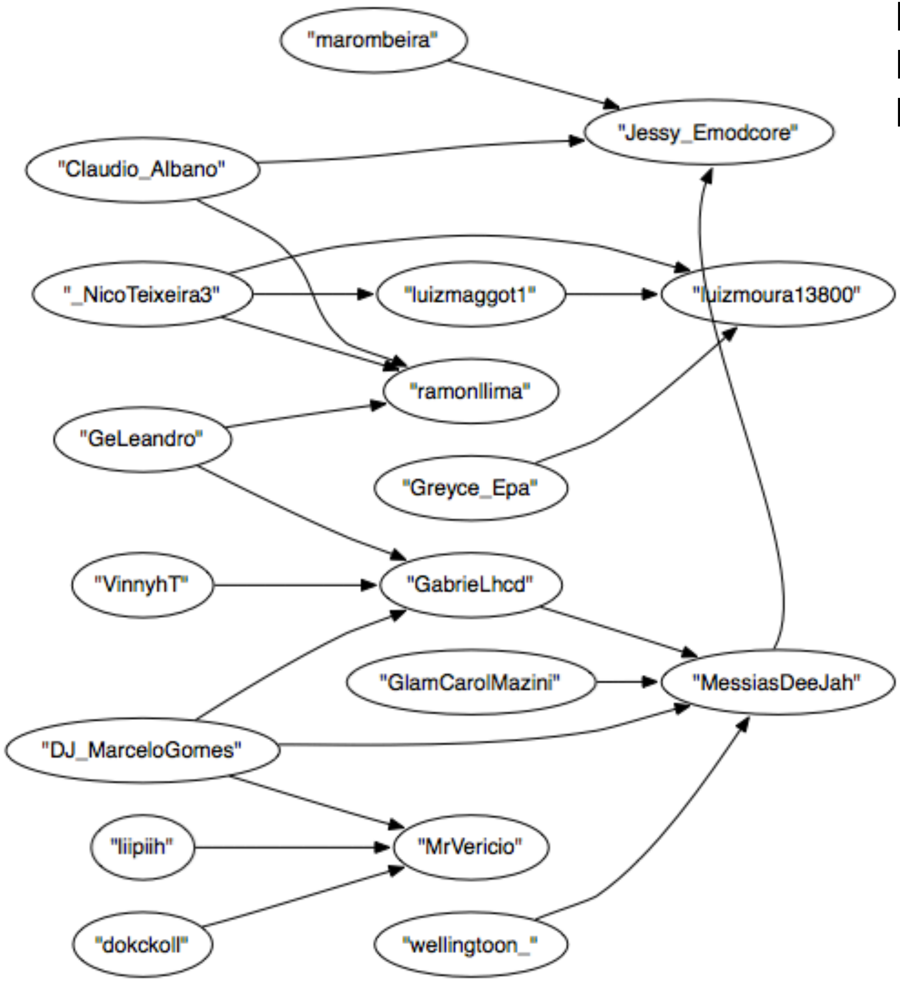
Free2BurnMusic: "#Nowplaying Janet Jackson - Hot 100 1990 http://free2burn.com/index.php #Music #IFollowBack #Music"

1 second later

Free2Burn: "#Nowplaying Janet Jackson - Hot 100 1990 http://free2burn.com/index.php #Music #IFollowBack #Music"

Bombe cluster

- High transfer entropy among users with most followers



BOMBE O SEU TWITTER, COM MILHARES DE NOVOS FOLLOWERS, ATRAVES DO SITE: http://????????#QueroSeguidores NNN

Google Translate:
Pump up your Twitter, get thousands of new followers, link to this site: http://?????? #IWantFollowers NNN

Links and numbers changing over time,
Most users re-posted many times.

Tweeted over 50,000 times.

Two users with same TE

Marina Silva
@silva_marina Brasil
Sou professora de História. Fui candidata à Presidência da República pelo PV em 2010, ministra do Meio Ambiente(2003-2008) e senadora pelo Acre, de (1995-2011).
http://www.minhamarina.org.br

Total TE ≈ 0.025

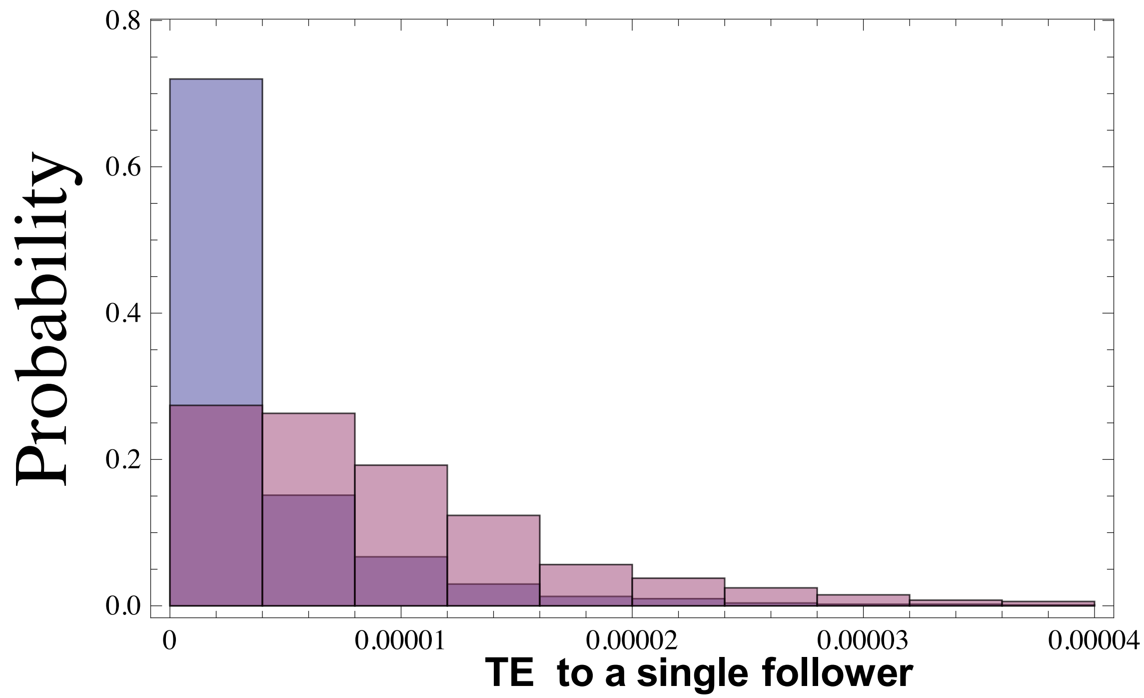
514,347 Followers

Soulja Boy (S.Beezy)
@souljaboy Atlanta, GA
*President of SODMG: Producer/Artist/Gamer/Student signed to Collipark Music/Interscope Records living a dream... $$$ * #SWAG #energy*
https://plus.google.com/116381176537835440497/

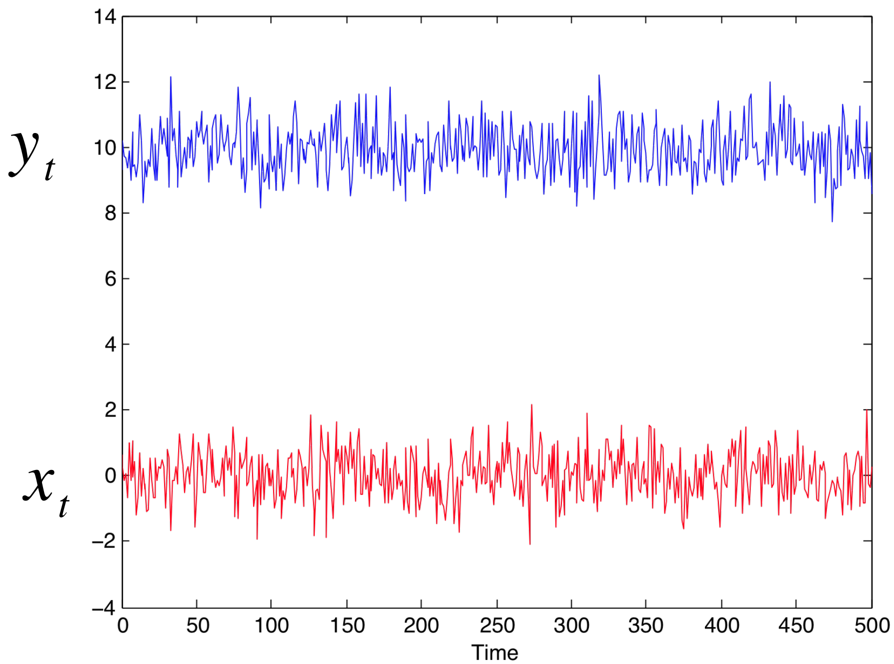
Total TE ≈ 0.025

3,110,453 Followers

Data taken just before the Brazilian presidential elections, for which Marina was a top contender.
Soulja Boy has many more followers, but most are only weakly influenced.



Granger Causality

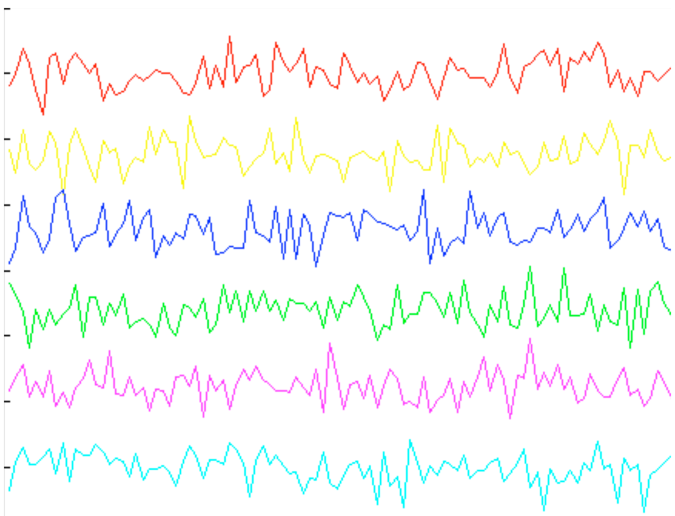


$$x_{t+1} \approx \sum_{j=1}^{p} A_j x_{t-j} + \sum_{j=1}^{l} B_j y_{t-j}$$

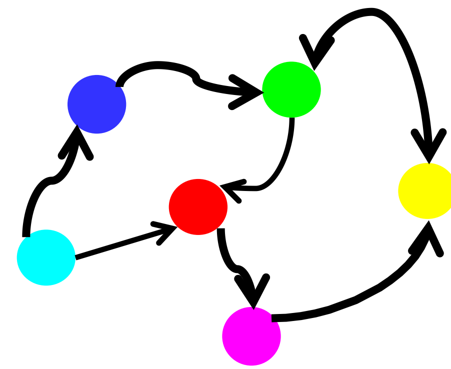
- Time series might represent
  - #of tweets by a user in a given time interval (e.g., per day)
  - # of certain hashtag mentions
  - etc

Straightforward Approach

- Calculate all pair-wise influence between the time series



$$x_i^{t+1} = \sum_{j=1}^{l} \beta_{ji} x_j^{t,Lagged}$$



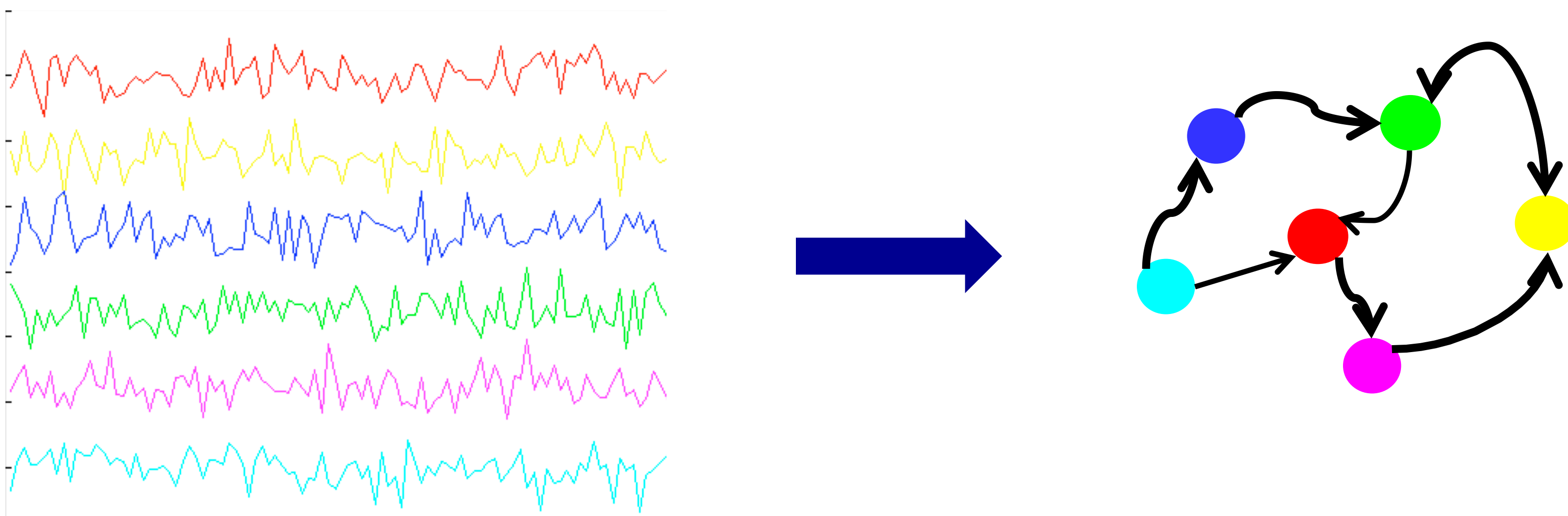
Problem: The learned influence network will be generally very dense

Granger Graphical Models

- Combining Granger-causality and variable selection

$$\hat{\beta}^{a,\lambda} = \operatorname{argmin} \sum_{i=1}^{n} \left\| x_i^a - X_{-a,i} \cdot \beta \right\|_2 + \lambda \left\| \beta \right\|_1$$

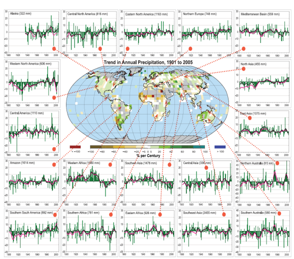
Sparsity term

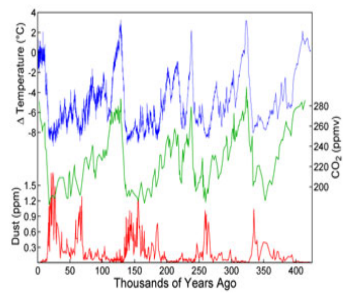


- Results in sparser (simpler! network)

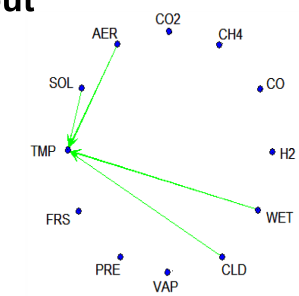
Granger Graphical Models

- Climate time-series analysis for climate-forcing agents
  [Lozano et.al., KDD'09]

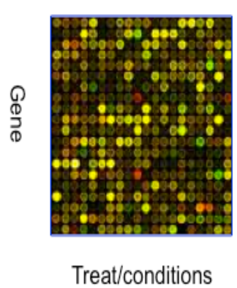


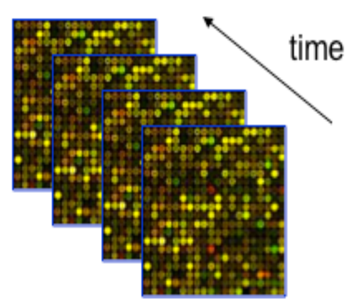


Output



- Time-series microarray analysis for regulatory dependencies
  [Liu et. el, ISMB'09]





Output

Uncovering hidden influence networks

- **Information transfer** (Schrieber, 2000)
  - How much is our uncertainty about user X's future activity reduced by knowing about Y's past activity?

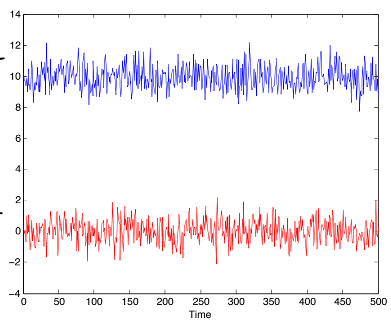
$$IT_{Y \to X} = \underbrace{H(X^{Future} | X^{Past})}_{\text{Uncertainty about X}} - \underbrace{H(X^{Future} | X^{Past}, Y^{Past})}_{\text{Uncertainty about X, if you know Y}}$$

    – *Arbitrary signals/relationships; hard to evaluate*

- **Granger Causality** (Granger, 1969)

Y

X



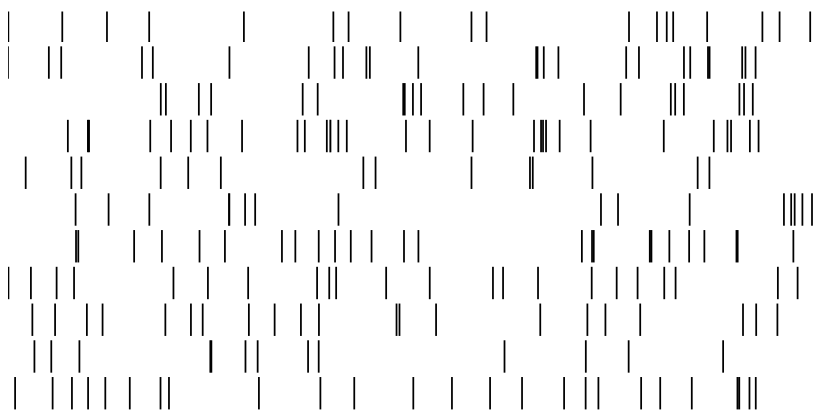
$$(M1) \quad X_{t+1} = \sum_{j=1}^{p} A_j X_{t-p}$$

$$(M2) \quad X_{t+1} = \sum_{j=1}^{p} A_j X_{t-p} + \sum_{k=1}^{m} B_j Y_{t-k}$$

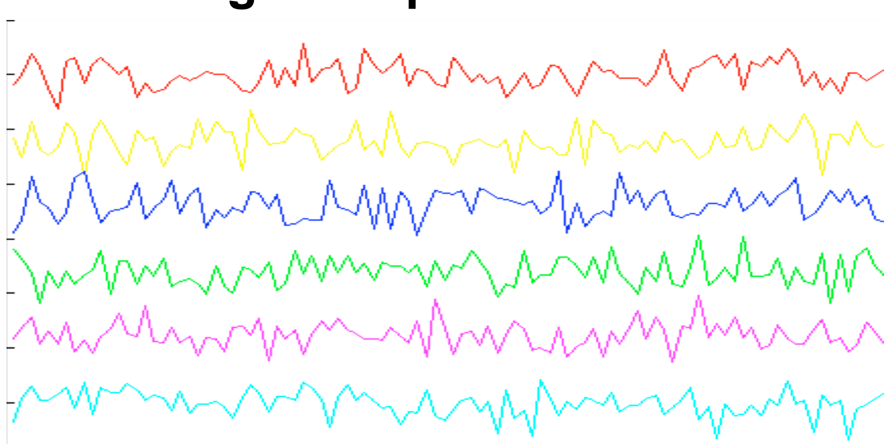
  - **Y** is "Granger-causal" to **X** if (M2) is a better predictor than (M1)
    – *More efficient but assumes linearity; real-valued signals only*

Summary

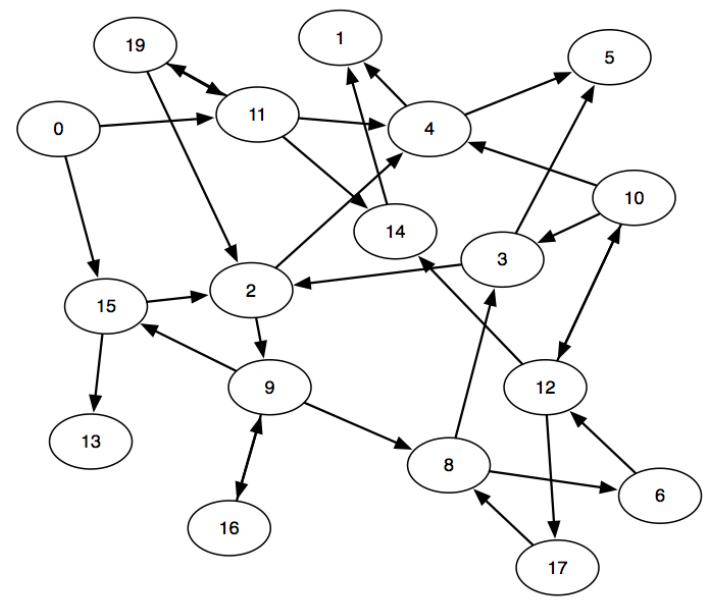
Information Transfer from Activity Timing



Granger Graphical Models



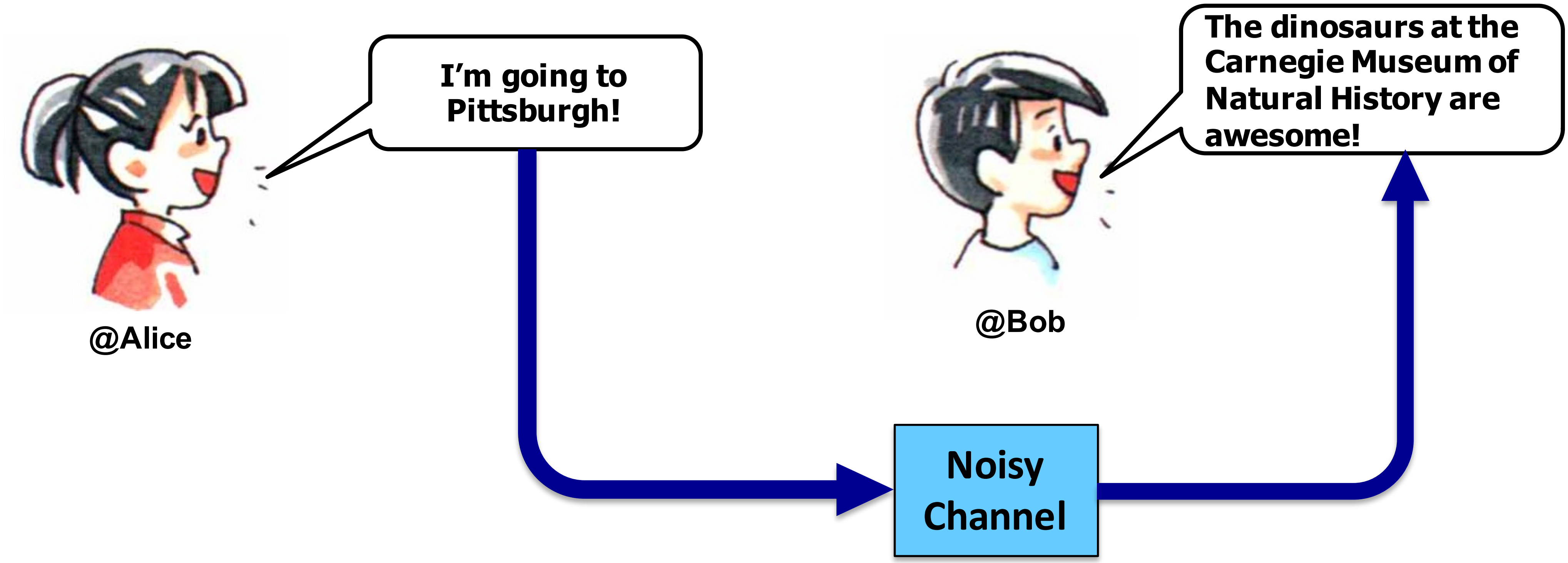
Arbitrary signals/representation, but hard to evaluate



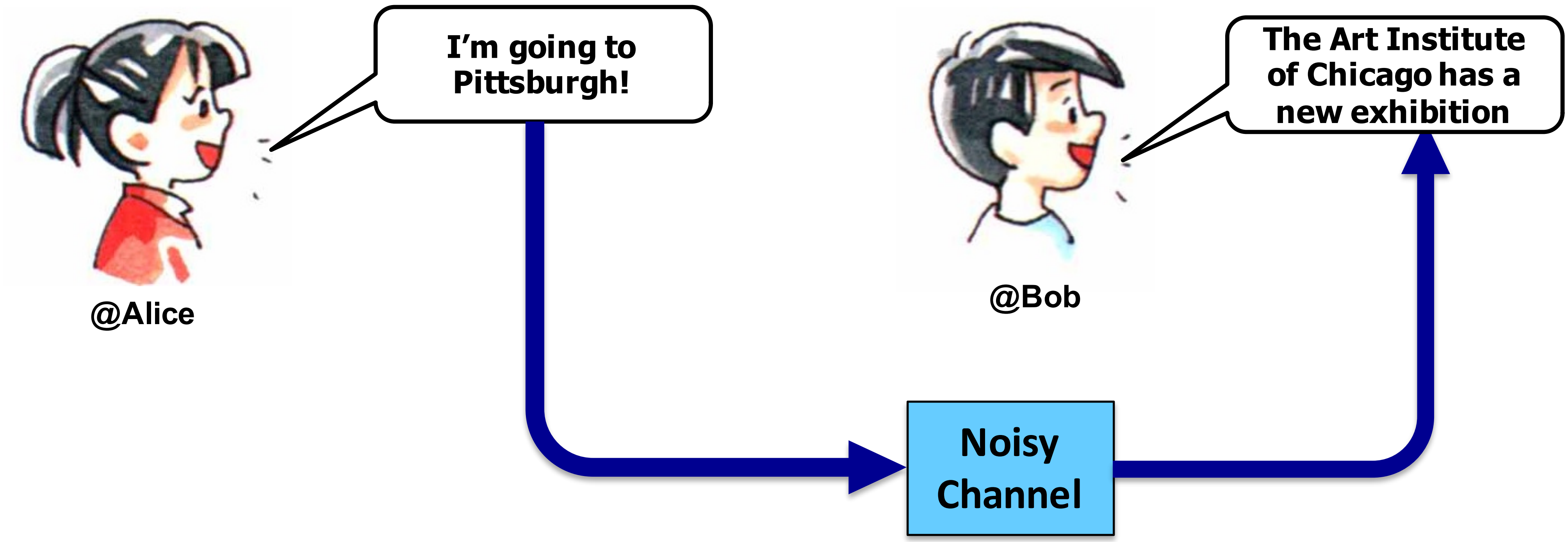
More efficient but assumes linearity; real-valued signals only

Inferring Social Influence from Content

Information in human speech



Information in human speech



$$I(A:B)$$

How much information is communicated?

Information in human speech

- Mutual information between Alice and Bob's statements:

$$I(A:B) = \sum_{A,B} P(A,B) \log \frac{P(A,B)}{P(A)P(B)}$$

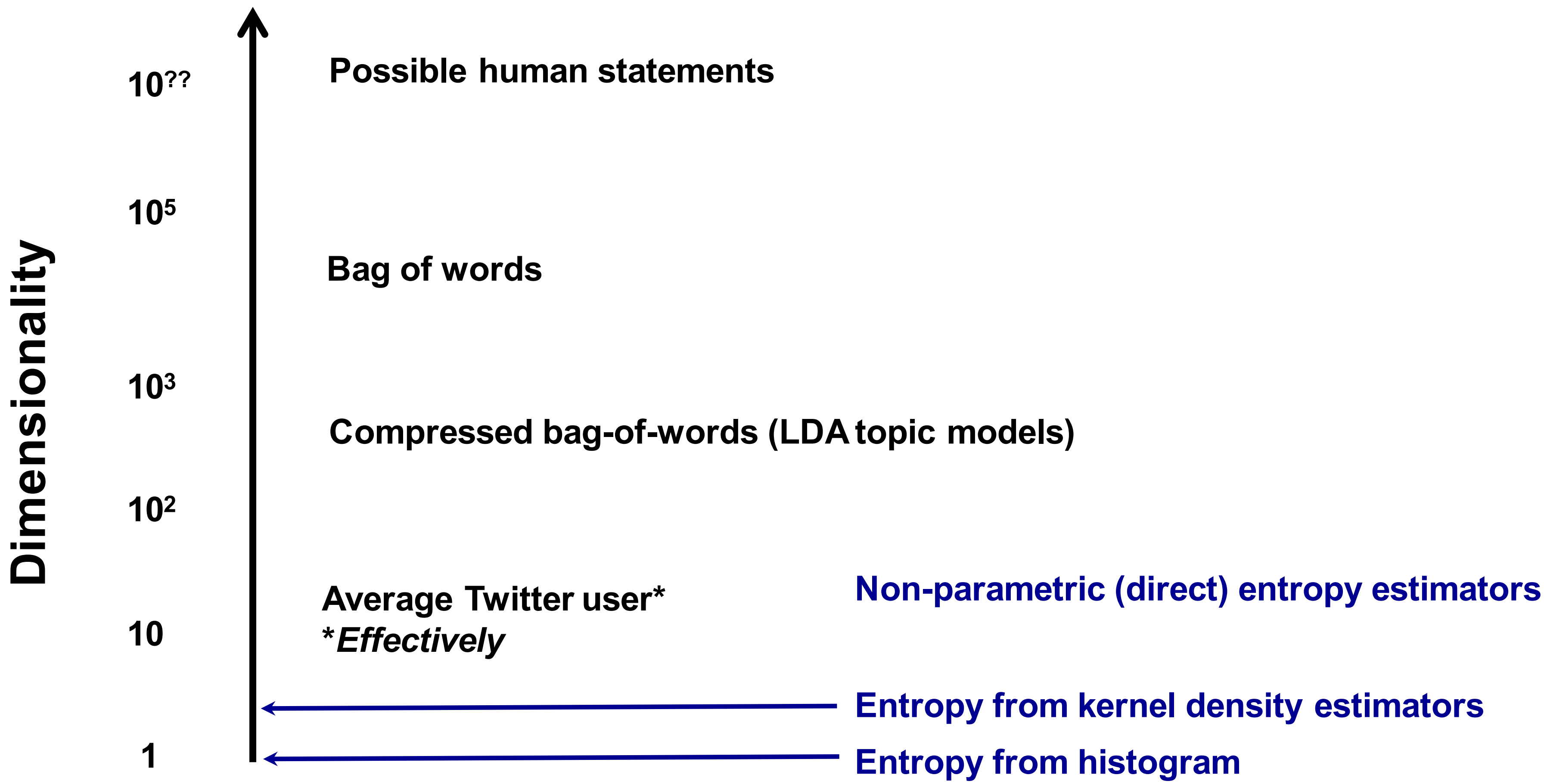
Sum over all possible statements!

- Includes such hard to quantify probabilities as:

Pr(Alice says **"I'm going to Pittsburgh**", then Bob says "**Dinosaurs are awesome**"**)**

- And, this is different for each pair of people!

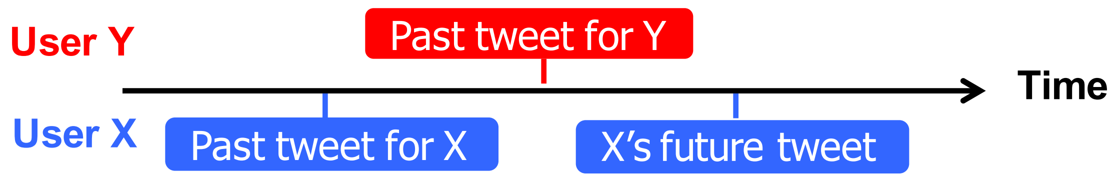
You’re so 10 dimensional



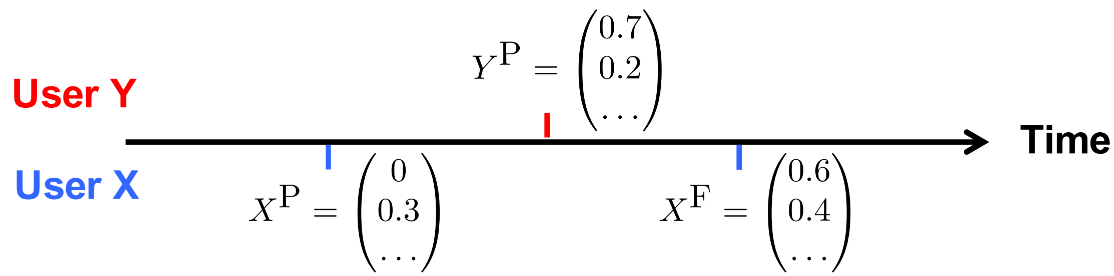

I'm going to Pittsburgh!
The dinosaurs at the Carnegie Museum of Natural History are awesome!
The Art Institute of Chicago has a new exhibition
(yesterday)
LACMA…
(last week)

Overview

- N samples of tweet exchanges



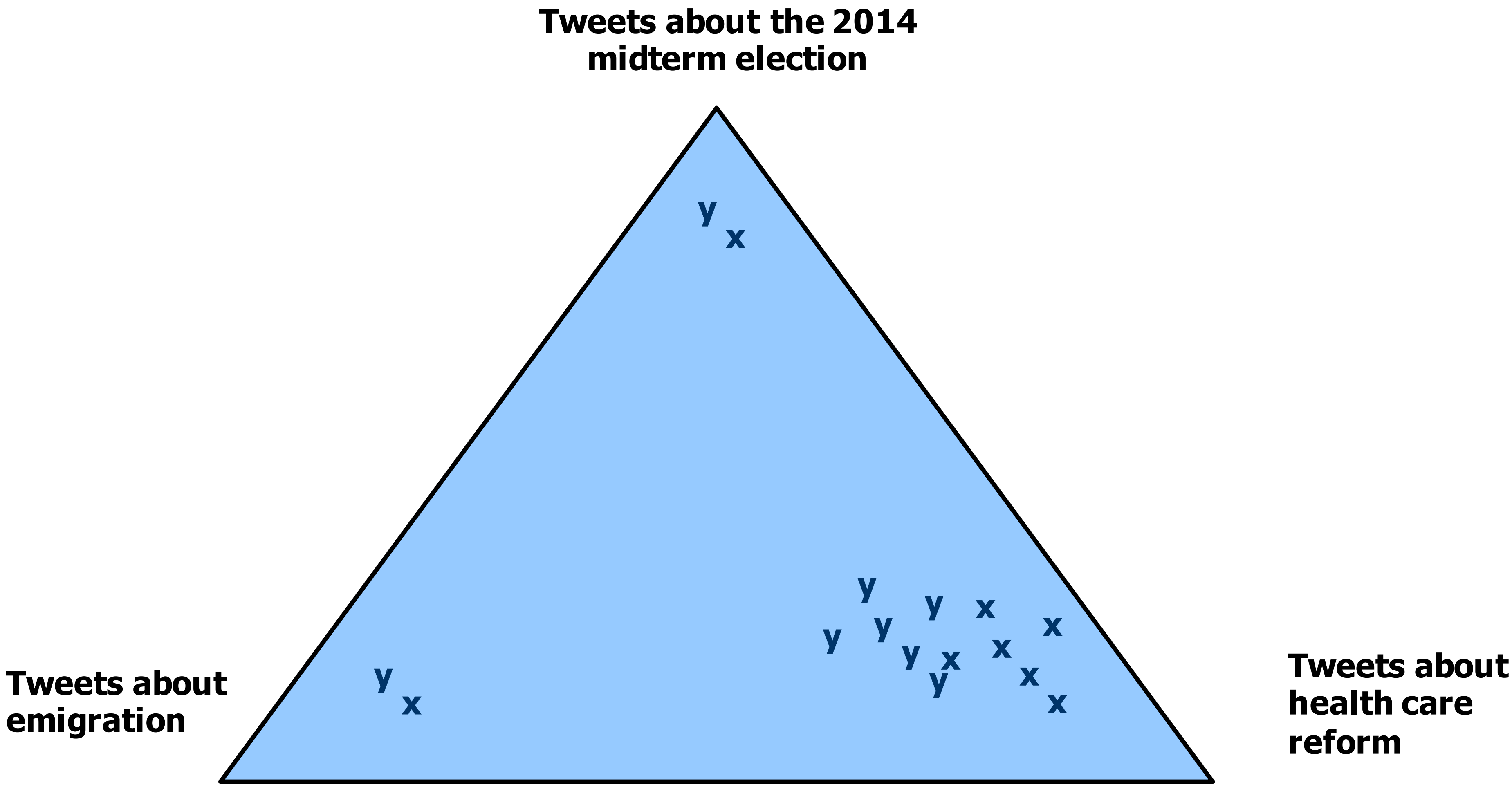
- Convert to an abstract representation



- Estimate transfer entropy: measure of Y's predictivity of X

$$TE_{Y \to X} = \hat{I}(X^F : Y^P | X^P)$$

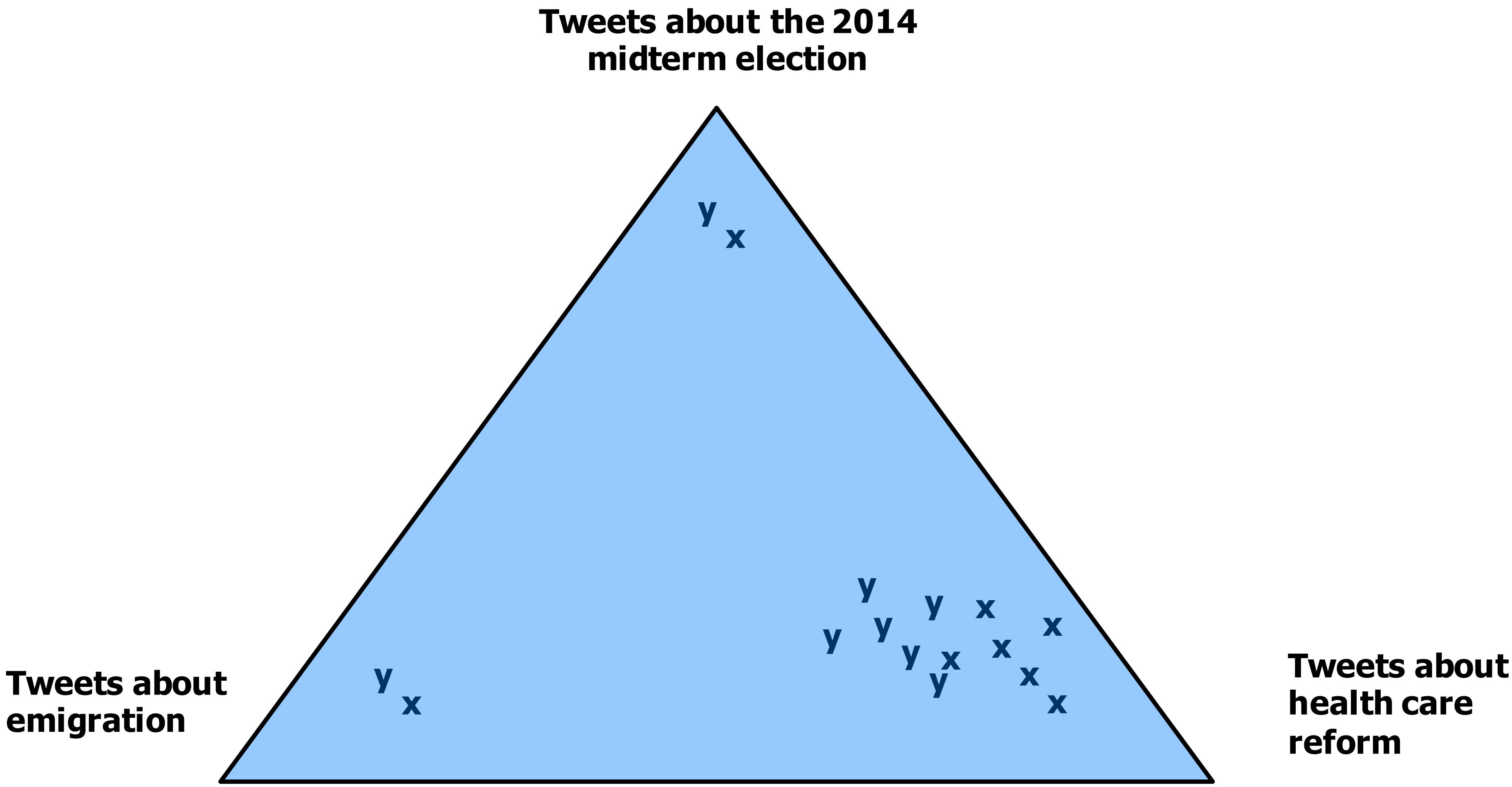
Predictability in content space



High transfer entropy : x's tweet was
more predictable from y's recent tweet
than from his own past tweets

Low Transfer Entropy – X is already predictable

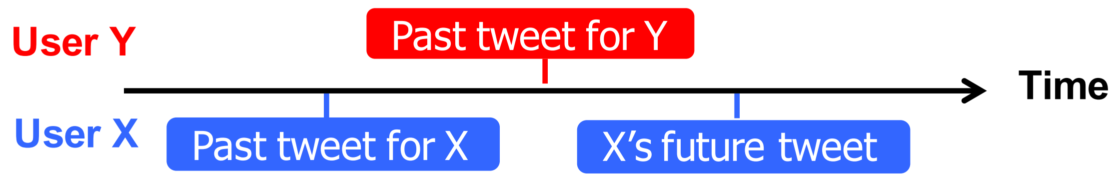
Predictability in content space



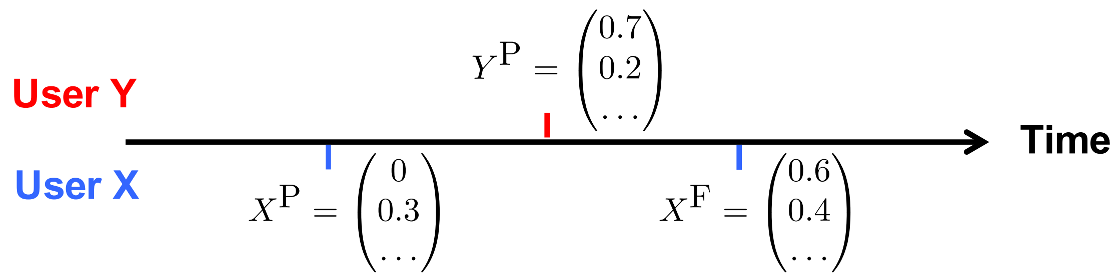
High transfer entropy : x's tweet was more predictable from y's recent tweet than from his own past tweets

Overview

- N samples of tweet exchanges



- **Convert** to an abstract representation



- **Estimate** transfer entropy: measure of Y’s predictivity of X

$$TE_{Y \to X} = \hat{I}(X^F : Y^P | X^P)$$

Convert to an abstract representation

Easiest: we'll use LDA topic model vectors from *gensim*. Best?

$$\longrightarrow \begin{array}{cl} \left(\begin{array}{c} 0.01 \\ 0.32 \\ 0.61 \\ 0.04 \\ \ldots \end{array}\right) & \begin{array}{l} \text{Music} \\ \text{Religion} \\ \text{Aviation} \\ \text{Livestock} \\ \ldots \end{array} \end{array}$$

Estimate transfer entropy

$$X^{\mathrm{P}}, Y^{\mathrm{P}}, X^{\mathrm{F}} = \begin{pmatrix} 0.6 \\ 0.4 \\ \ldots \end{pmatrix}, \begin{pmatrix} 0.1 \\ 0.3 \\ \ldots \end{pmatrix}, \begin{pmatrix} 0.2 \\ 0.8 \\ \ldots \end{pmatrix} \longrightarrow TE_{Y \to X}$$

~100 samples of ~100-dim topic vectors!

(luckily, most users' activity is effectively low-d)

Non-parametric entropy estimators

- No binning of data
- No estimating probability density
- Nice convergence properties

Twitter study

- 1 month of tweets
- ~2k users, snowball sampling, constrained to Middle East
- 768k tweets

- PREPROCESSING:
  - **No RTs**
  - [a-zA-Z] only, lowercased
  - No punctuation
  - No stop words

- Calculate transfer entropy for all ordered pairs of users

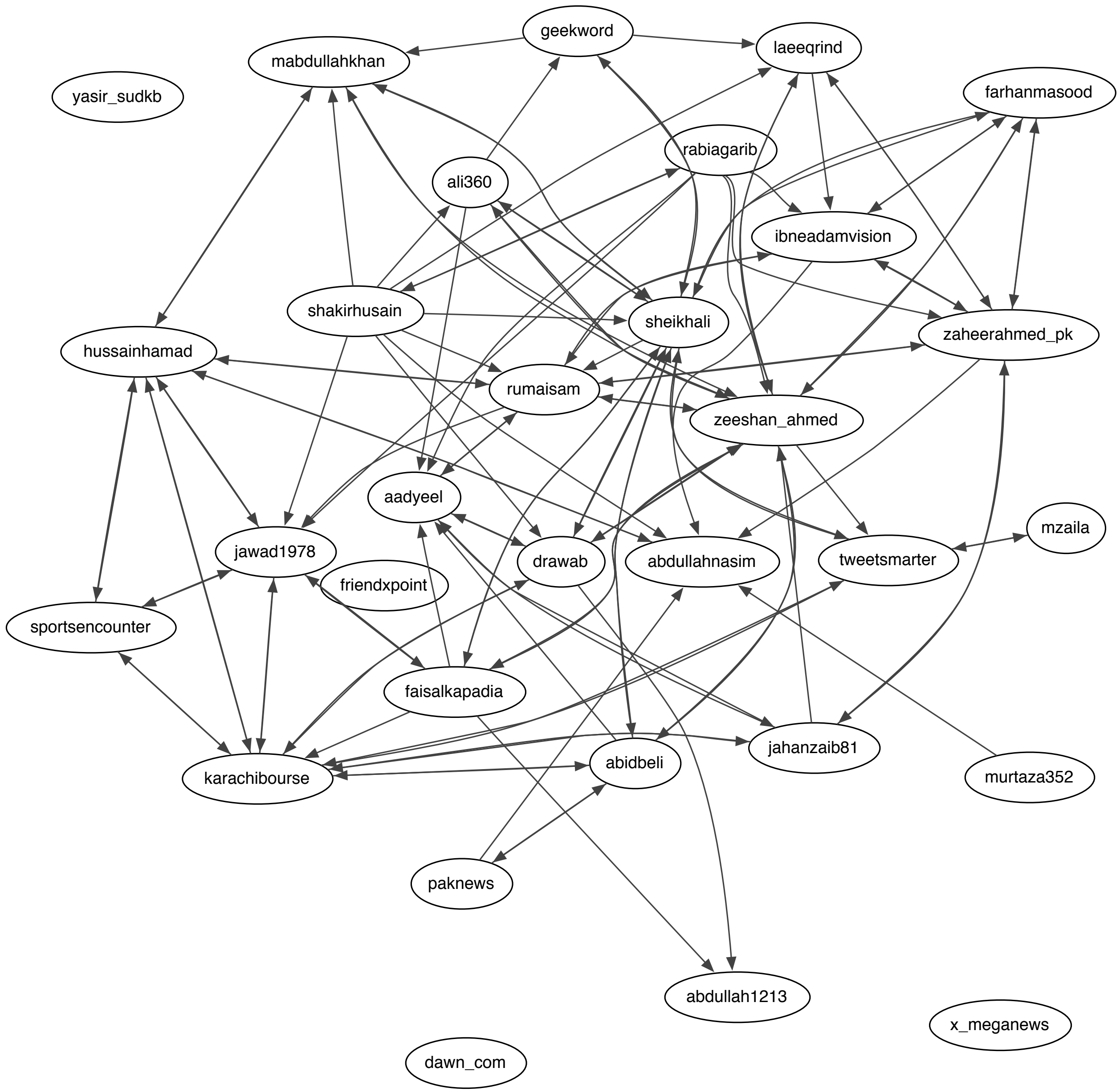
Histogram of transfer entropy

Pairs

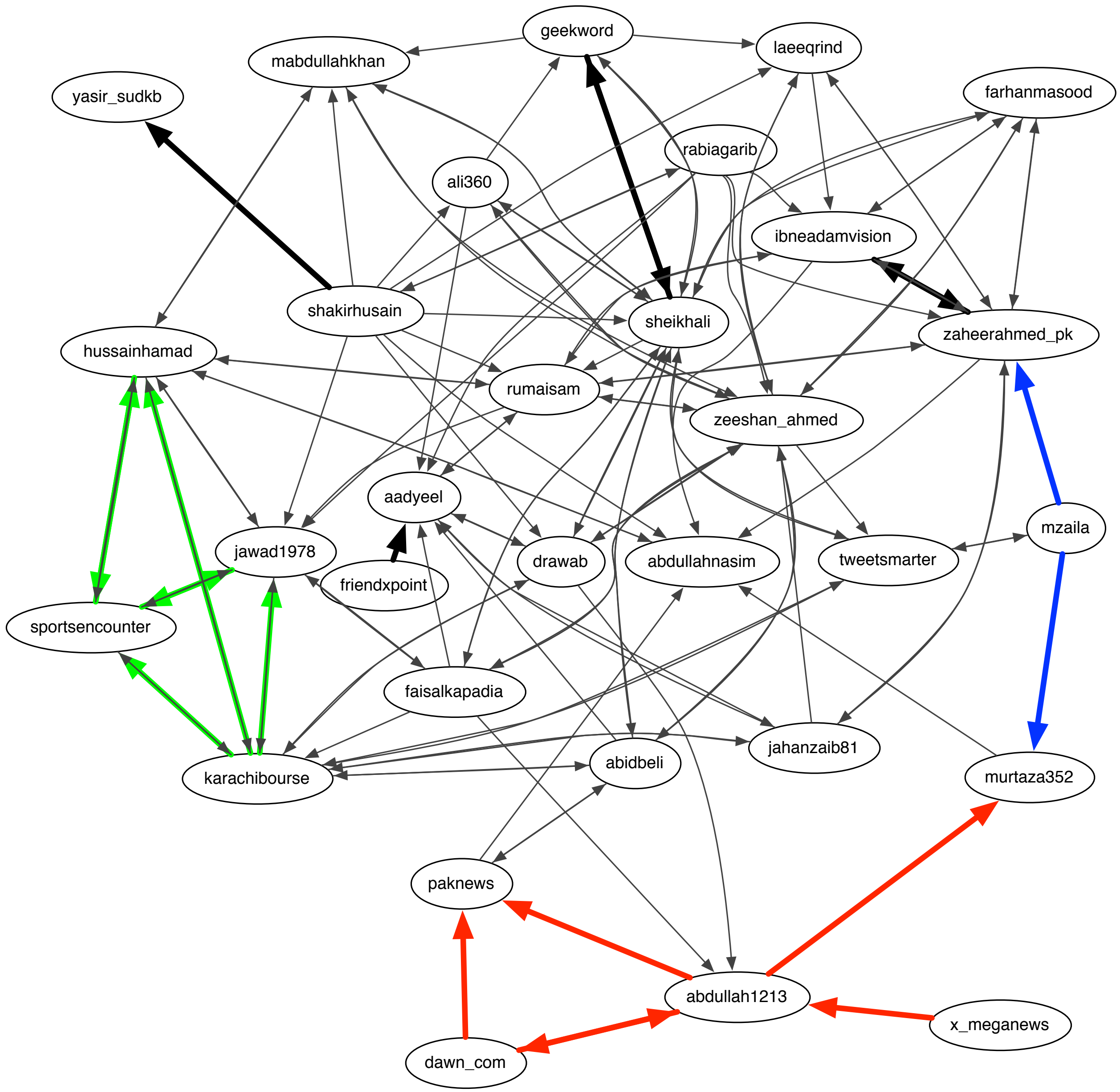
Very high transfer entropy
pairs!

Transfer
Entropy

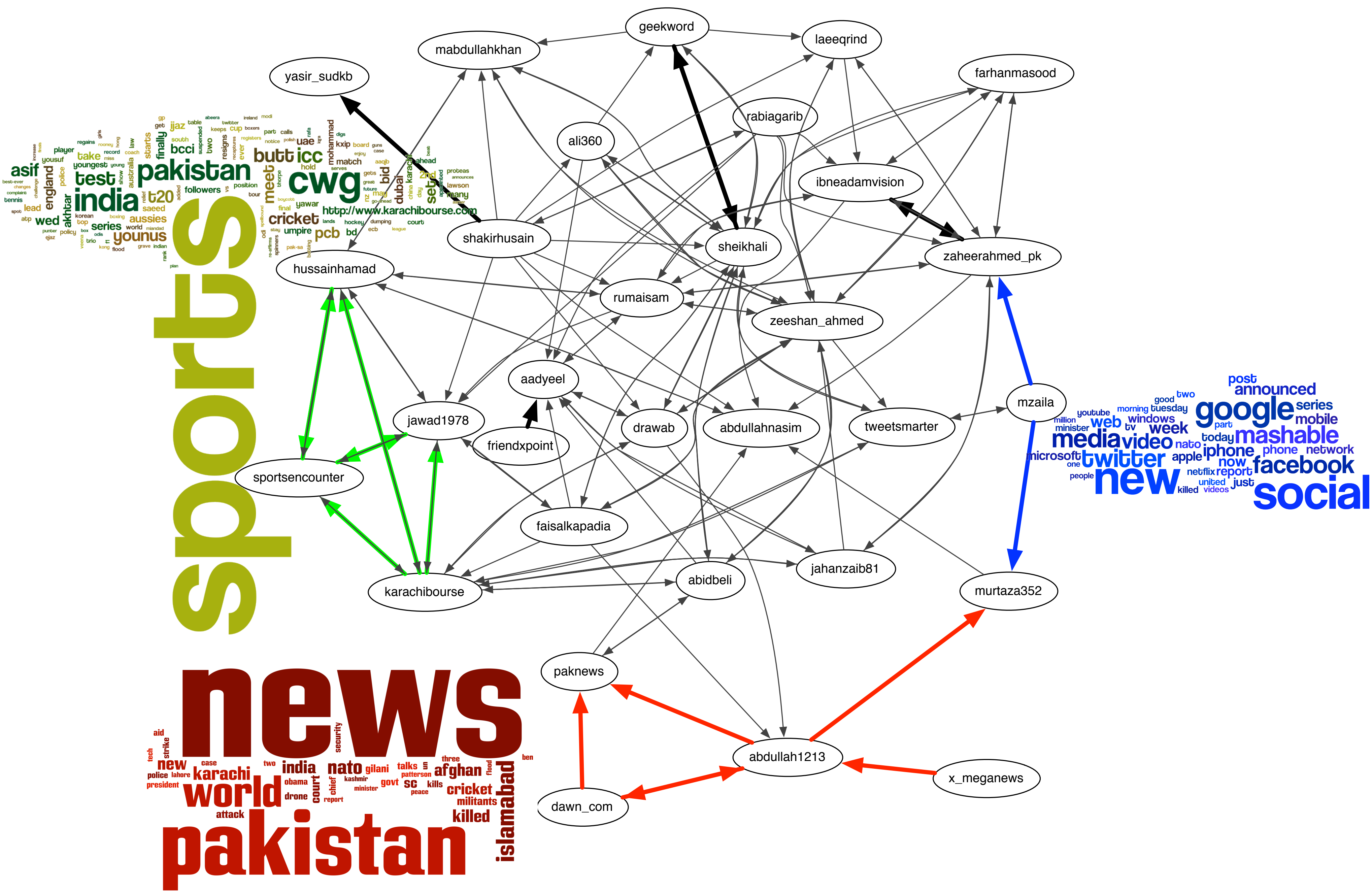
“Friend-follower” network



Transfer entropy network

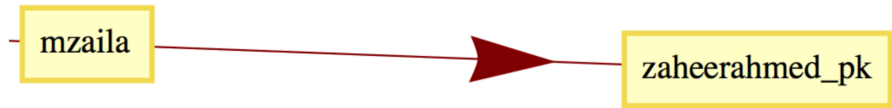


Transfer entropy network



-No follows
-No retweets
-Random order leads to bi-directed transfer

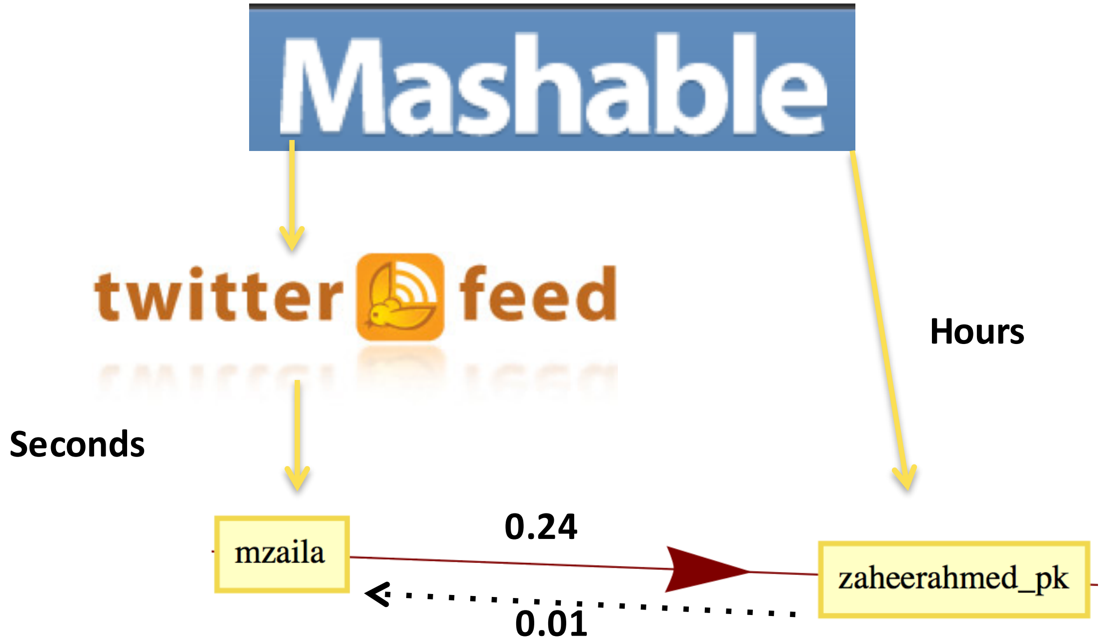
geekword: #Skype for #Windows gets deep rooted #Facebook Integration http://bit.ly/cb7UOj #SocialNetwork
sheikhali: #Skype for #Windows gets deep rooted #Facebook Integration http://bit.ly/cb7UOj #SocialNetwork
sheikhali: @l3v5y nice one
geekword: #Windows Phone 7 to get copy/paste feature in early 2011 http://bit.ly/a9AfF5 #Wp7 #Microsoft #gadgets
sheikhali: #Windows Phone 7 to get copy/paste feature in early 2011 http://bit.ly/a9AfF5 #Wp7 #Microsoft #gadgets
geekword: #Windows Phone 7 makes a guest appearance on #HTC #HD2 http://bit.ly/aUJmJp #WP7
sheikhali: #Windows Phone 7 makes a guest appearance on #HTC #HD2 http://bit.ly/aUJmJp #WP7
geekword: Where to watch #Apple's Back to the Mac event streamed live http://goo.gl/fb/843kl #gadgets #newsreviews #macbookair
sheikhali: How to watch live streaming of #Apple's Back to the #Mac Event http://bit.ly/bGJ4w2 #gadgets #Macbook
sheikhali: @geekword trending post: #Ultrasn0w #iOS 4.1 #unlock for #iPhone 3G(S) will go live two days after the iOS 4.2 release http://bit.ly/9QKcNB
geekword: #PwnageTool 4.1 unleashed brings iOS 4.1/3.2.2 #jailbreak for your #iDevice http://bit.ly/cn50Qu #Apple #jbiPhone
sheikhali: #PwnageTool 4.1 unleashed brings iOS 4.1/3.2.2 #jailbreak for your #iDevice http://bit.ly/cn50Qu #Apple #jbiPhone
geekword: @tweetmeme How to watch live streaming of #Apple's Back to the #Mac Event http://bit.ly/bGJ4w2 #gadgets #Macbook
sheikhali: @tweetmeme How to watch live streaming of #Apple's Back to the #Mac Event http://bit.ly/bGJ4w2 #gadgets #Macbook
geekword: #Guide to #jailbreak iOS 4.1 using #PwnageTool 4.1 http://bit.ly/bz6dv8 #jbiPhone #Howto
sheikhali: #Guide to #jailbreak iOS 4.1 using #PwnageTool 4.1 http://bit.ly/bz6dv8 #jbiPhone #Howto
geekword: @tweetmeme #Guide to #jailbreak iOS 4.1 using #PwnageTool 4.1 http://bit.ly/bz6dv8 #jbiPhone #Howto
sheikhali: @tweetmeme #Guide to #jailbreak iOS 4.1 using #PwnageTool 4.1 http://bit.ly/bz6dv8 #jbiPhone #Howto



User	Tweet
zah	KARACHI, Pakistan, Oct. 12 (UPI) – Intelligence agencies in Pakistan are warning of terrorist atta... http://bit.ly/bscYoX #news #Pakistan
mza	Is Mobile Video Chat Ready for Business Use?: Matthew Latkiewicz works at Zendesk.com, creators of web-based custo... http://bit.ly/cAx3Ob
zah	Matthew Latkiewicz works at Zendesk.com, creators of web-based customer support software. He writes for... http://bit.ly/bkuWCV #technology
zah	Man-made causes cited for Pakistan floods: ISLAMABAD, Pakistan, Oct. 14 (UPI) – Deforestation ... http://bit.ly/92afA0 #pkfloods #Pakistan
mza	Google Shares Jump 7% on Impressive Earnings: Google has posted its latest earnings report, and early indications ... http://bit.ly/9oi4zr
zah	Google has posted its latest earnings report, and early indications suggest that investors are more tha... http://bit.ly/cyT35p #technology

No following
No mentions
No RT
Different URL
Different Hash
Different wording

LTE puts exchanges about same story higher with probability 0.68



Asymmetric:
Temporally, only one order occurs (mza then zah)
It's *predictable* but is it *causal?*

LTE	User	Tweet
2.65	zah	KARACHI, Pakistan, Oct. 12 (UPI) – Intelligence agencies in Pakistan are warning of terrorist atta... http://bit.ly/bscYoX #news #Pakistan
	mza	Is Mobile Video Chat Ready for Business Use?: Matthew Latkiewicz works at Zendesk.com, creators of web-based custo... http://bit.ly/cAx3Ob
	zah	Matthew Latkiewicz works at Zendesk.com, creators of web-based customer support software. He writes for... http://bit.ly/bkuWCV #technology
2.53	zah	Man-made causes cited for Pakistan floods: ISLAM...

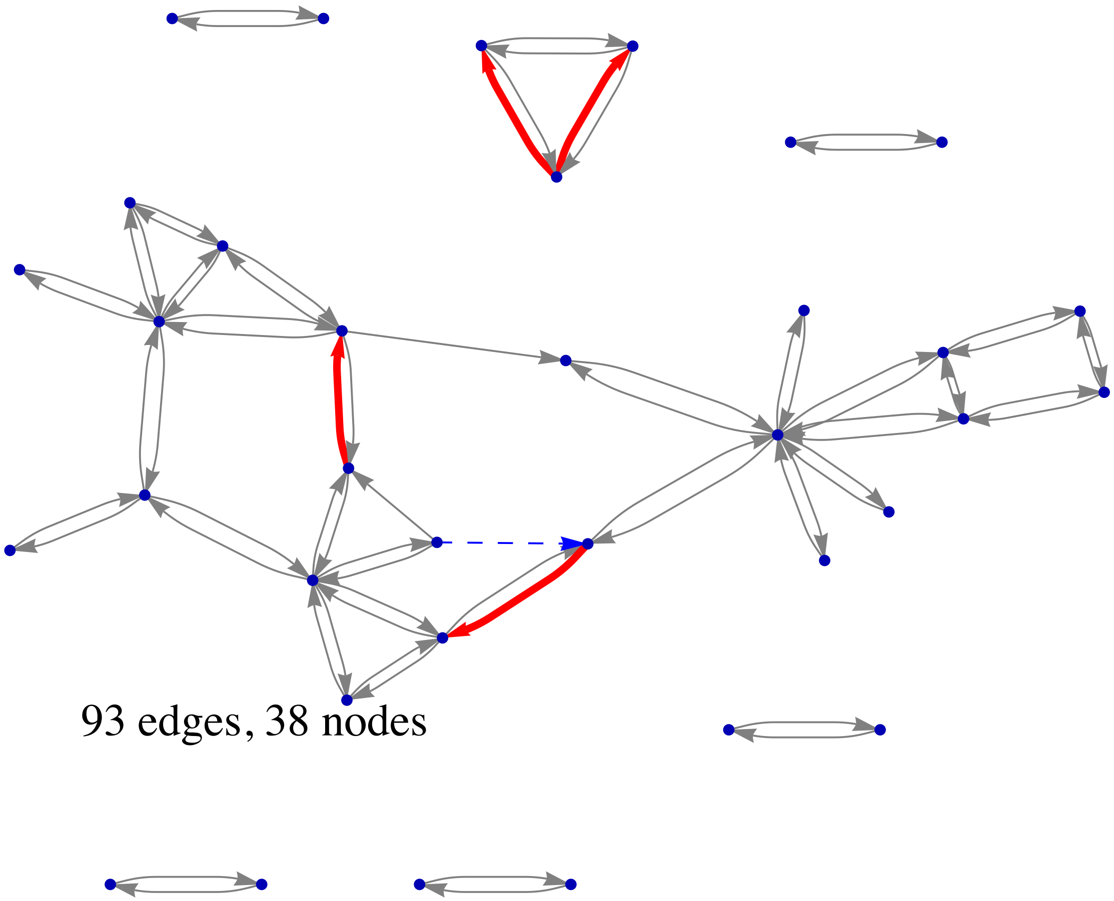
Social influence

Previous examples were *predictable* but not *social*

- Can we use mentions to check if we capture social behavior?

- We consider to a subset of users who use mutual mentions in conversation

Reconstructing mention graph



93 edges, 38 nodes

Top 4 edges according to transfer entropy are correct:

"tabankhamosh", "shahidsaeed", 0.110
"noy_shahar", "lihifarag", 0.0987
"enggandy", "fzzzkhan", 0.0976
"noy_shahar", "reutgolan", 0.0975

Metric:
Probability that a true edge has higher transfer entropy than a false edge

AUC = 0.648

Null model: **AUC = 0.5**

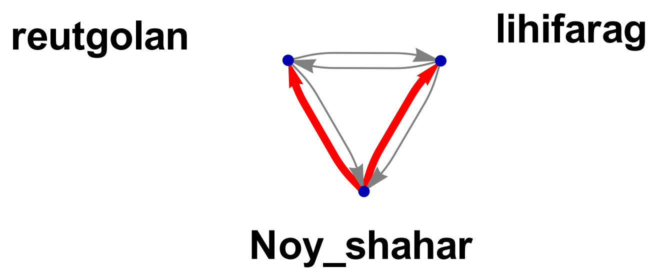
(w/ SE = 3.5%)

Top transfer entropy examples

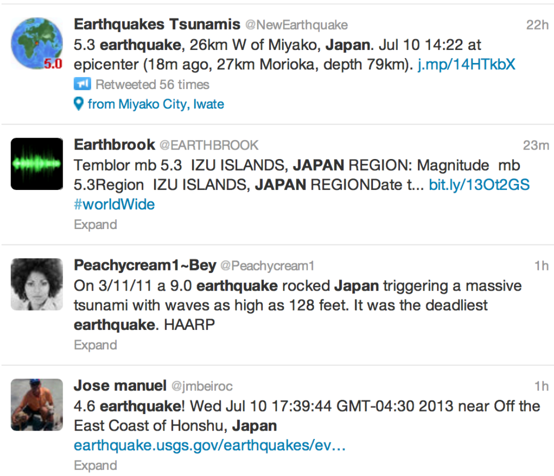
User	Tweet
sh	@ta tsalk to police officers. 6 prominent policemen of Op Cleanup have been killed in last 2 yrs. Still tolerating MQM
ta	@sh I meant the "participation" of the hijacked public was a function of fear perp by Talibs. Same thing here. ppl don't want 2 die
sh	@ta what does it serve them?More pathetic f*tards snatching their mobiles and wallets? Small-crime is engrained in MQM structure
ta	@sh re: "no soul n honor"... well I think MQM zia's creation to puncture the Sindh Nationalist cause. ISI _will_ slap its b*

Top transfer entropy examples

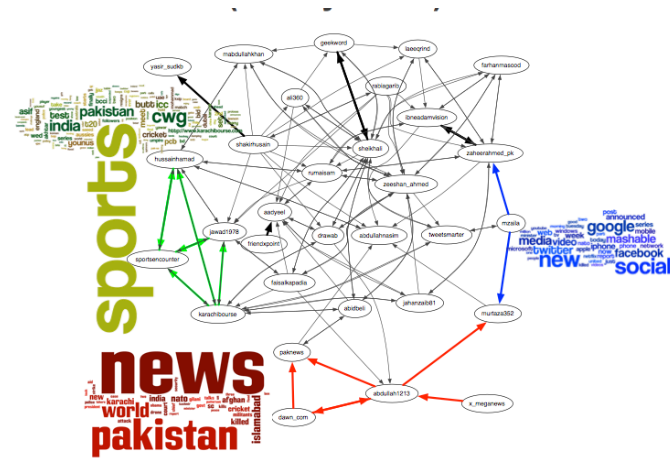
Tri-lingual friends



re	queremos unaa fotooooo deee @celeb1 y @celeb2
li	QUIERO UNA FOTO DE @celeb1 & @celeb2
no	@celeb2 nico .. please que la segunda imagen sera de vos con @celeb1
re	duele tanto decir ALGO ?
li	@celeb2 nico porfi saca una foto con emi :(
re	@No [Hebrew characters]
no	@Li @Re [Hebrew characters]
no	@re twiitcam baby, yes o no?!
re	@No yesssss, and my brother will be theirr !! hahah , your sweet
no	@Re jaja! very good sister! :)



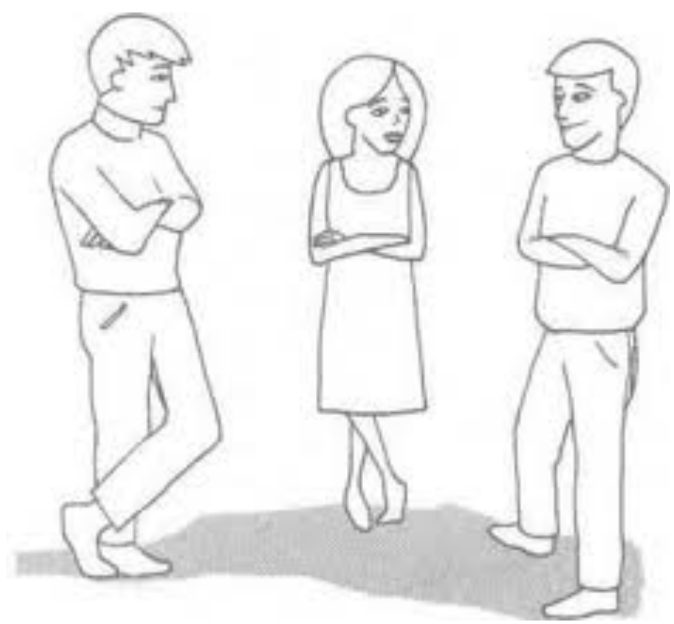
Summary



- Model-free approach to text-based analysis of social interactions
  - Grounded in Information Theory
  - Go beyond *followers, RT, #hash, URL.*
  - *Agnostic to representation (content, stylistic features, etc)*
  - *Can account for confounders by proper conditioning*
- Challenges and future work
- Better and/or different representation for text
- Better estimators for entropic measures

Stylistic Influence in Dialogues

Behavioral mirroring



Coordination in communication

- Communication Accommodation Theory:
  - When conversing, people non-consciously adapt to one another's communicative behaviors [Chartrand and Bargh, 1999]

Dimension	**Study**
Posture	Condon and Ogston, 1967
Head Nodding	Hale and Burgoon, 1984
Pause Length	Jaffe and Feldstein, 1970
Backchannels	White, 1984
Self-disclosure	Derlenga et al., 1973
Linguistic Style	Niederhoffer and Pennebaker, 2002
Linguistic Style **(Large Scale)**	Danescu-Niculescu-Mizil et al., 2011, 2012

Linguistic style coordination

- How things are said, rather what is said
- Example

  A: “What time are you available?”

  B: “Noon.”

Linguistic style coordination

- How things are said, rather what is said
- Example

  A: “What time are you available?”

  A: “At what time are you available?”

  B: “Noon.”

  B: “At noon.”

Linguistic style coordination

- How things are said, rather what is said
- Example

  A: “What time are you available?”

  A: “At what time are you available?”

  B: “Noon.”

  B: “At noon.”

- Quantified using function words (LIWC)
  - Reflect psychological processes [Chung & Pennebaker, 2007]
  - In this study: *articles*, *auxiliary verbs*, *conjunctions*, *adverbs*, *impersonal pronouns*, *personal pronouns*, *prepositions*, *quantifiers*

Function Words

- *Function words are processed rapidly and largely nonconsciously when people produce or comprehend language.* [Petten et al. 1991; Segalowitz et al., 2004]
- Linguistic Inquiry and Word Count(LIWC) [Pennebaker et al., 2007]

Category	Example
Personal Pronouns	I, them; her
Impersonal Pronouns	it, those
Articles	a, an, the
Auxiliary Verbs	am, will, have
Adverbs	very, really, quickly
Prepositions	to, with, above
Conjunctions	and, but, whereas
Quantifiers	few, many, much

Linguistic style coordination

Alice:	dfasdf to the dafgaf	(1,1)
Bob:	by dfa at dafsdf the dagfg	(1,1)
Alice:	dfasgfge of dfsd gaf dgevm	(1,0)
Bob:	drgt for dag fgfd	(1,0)
Alice:	dasf to dagftef an erfsadfa	(1,1)
Bob:	dfasd dag ad dagf dafs	(0,0)

……

……

red: prepositions blue: articles

Linguistic style coordination

Alice:	dfasdf to the dafgaf	(1,1)
Bob:	by dfa at dafsdf the dagfg	(1,1)
Alice:	dfasgfge of dfsd gaf dgevm	(1,0)
Bob:	drgt for dag fgfd	(1,0)
Alice:	dasf to dagftef an erfsadfa	(1,1)
Bob:	dfasd dag ad dagf dafs	(0,0)

- Coordination: Is Bob more likely to use a particular feature in his response, if Alice used that feature in her post?

$$\mathrm{Coord}(Bob \to Alice) = p(m_b = 1 | m_a = 1) - p(m_b = 1)$$

Prior results

- Observation of statistically significant coordination
  - Laboratory experiments [Pennebaker, 1999]
  - Large-scale experiments [Danescu-Niculescu-Mizil, 2012]
    - *Data from Supreme court transcripts & Wikipedia discussions*

- Stylistic coordination can be used to predict different behavioral outcomes
  - Relationship stability [Ireland, 2010]
  - Power relationship/social status [Danescu-Niculescu-Mizil, 2012]
  - Presidential debates & polling numbers [Romero 2015]

Alternative measure of stylistic coordination

- Given two users Alice and Bob and their corresponding feature sequence, we define stylistic coordination using (time-shifted) mutual information

m_A	m_B
0	0
1	0
0	1
0	0
1	0
…	…

$$\mathrm{Coord}(Bob \rightarrow Alice) = I(m_b^t : m_a^{t-1})$$

- For independent sequences the measure is identically zero

- Allows to consider possible confounders
  - E.g., length of utterances, conversation topic, etc

$$\mathrm{Coord}(Bob \rightarrow Alice) = CMI(m_b^t : m_a^{t-1} | Z)$$

Experiments

U.S. Supreme Court Oral arguments:

-50,000 verbal exchanges

-between **Justices** and **Lawyers**

Wikipedia Community of editors:

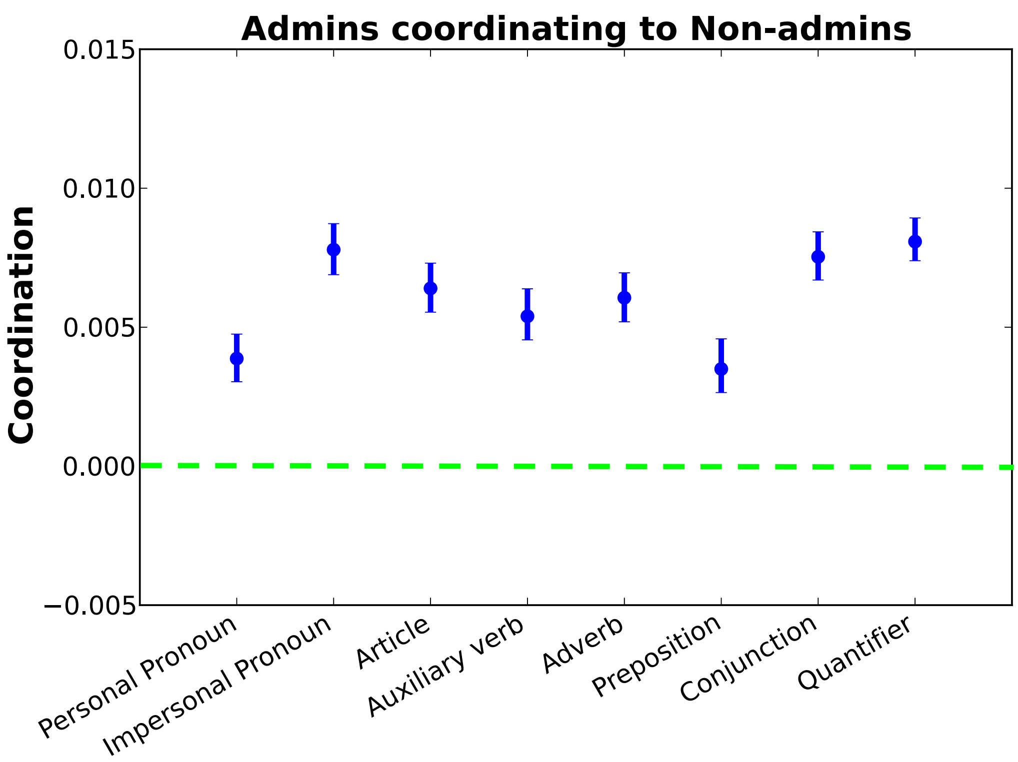
-240,000 conversational exchanges of discussions

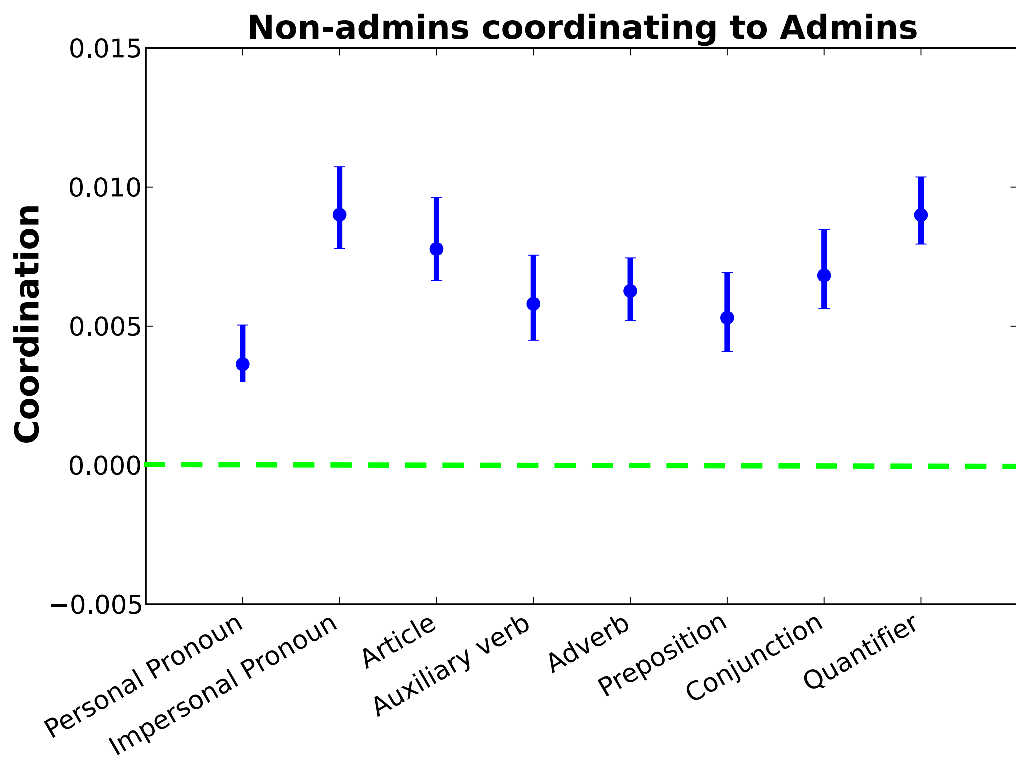
-users are either **admins** or **non-admins**

WIKIPEDIA
The Free Encyclopedia

Results

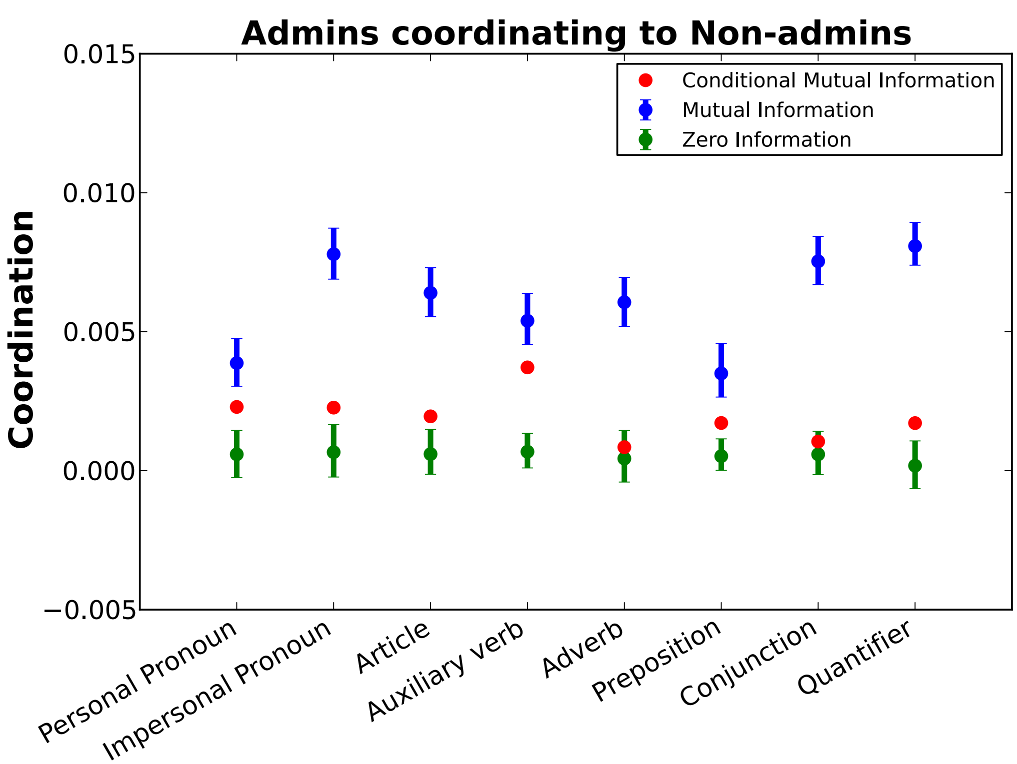
Wikipedia:

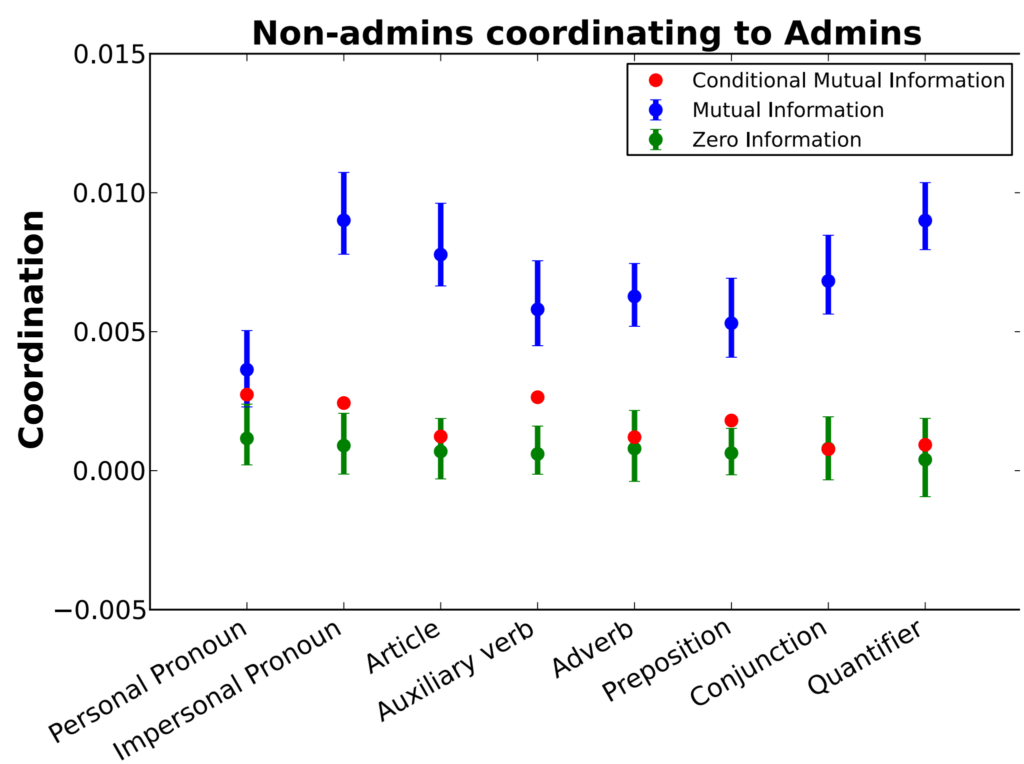




Results

Wikipedia: green error bars are obtained via shuffling the sequences

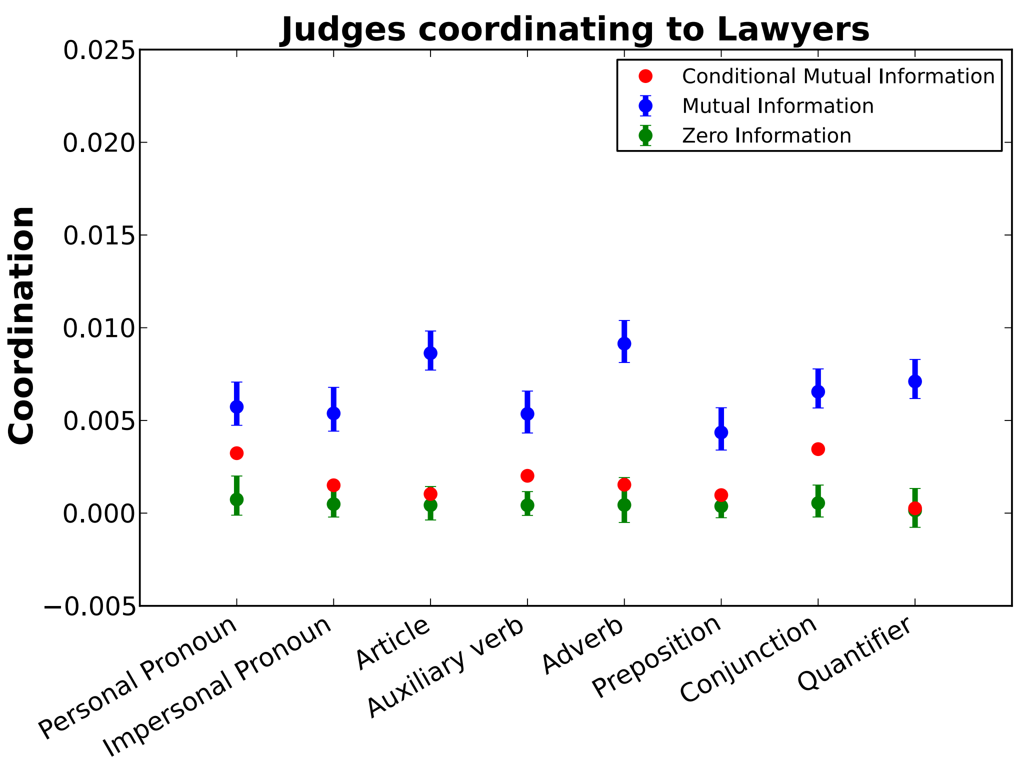


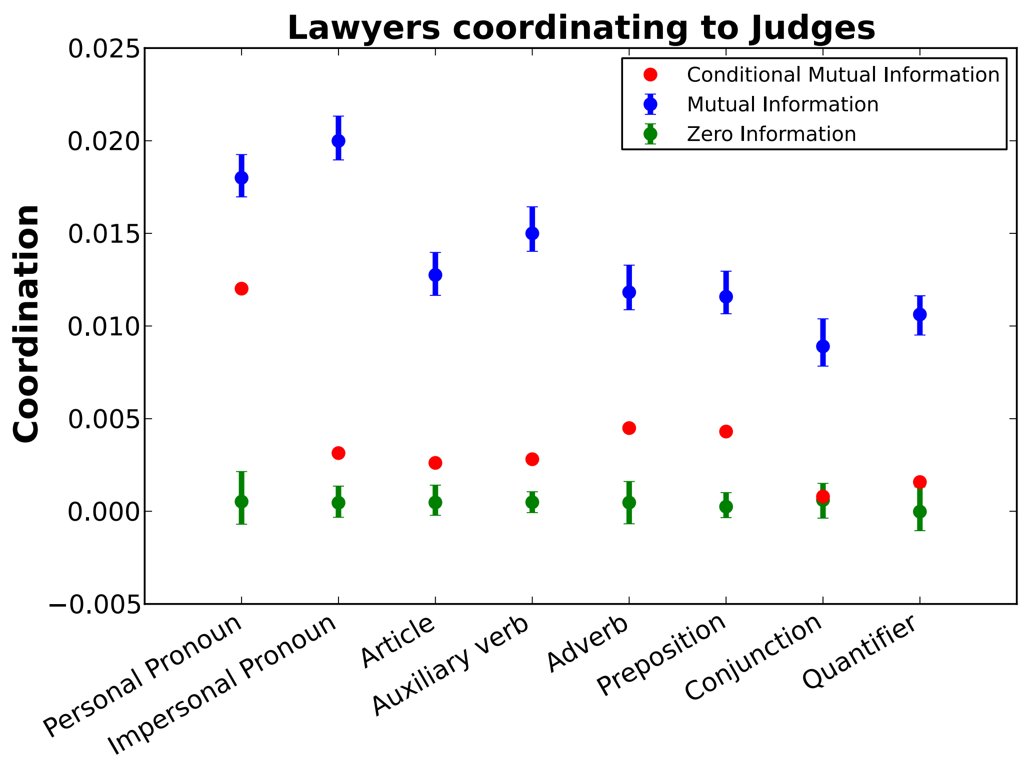


most "stylistic" coordination is "explained away" by length

Results

Supreme Court:

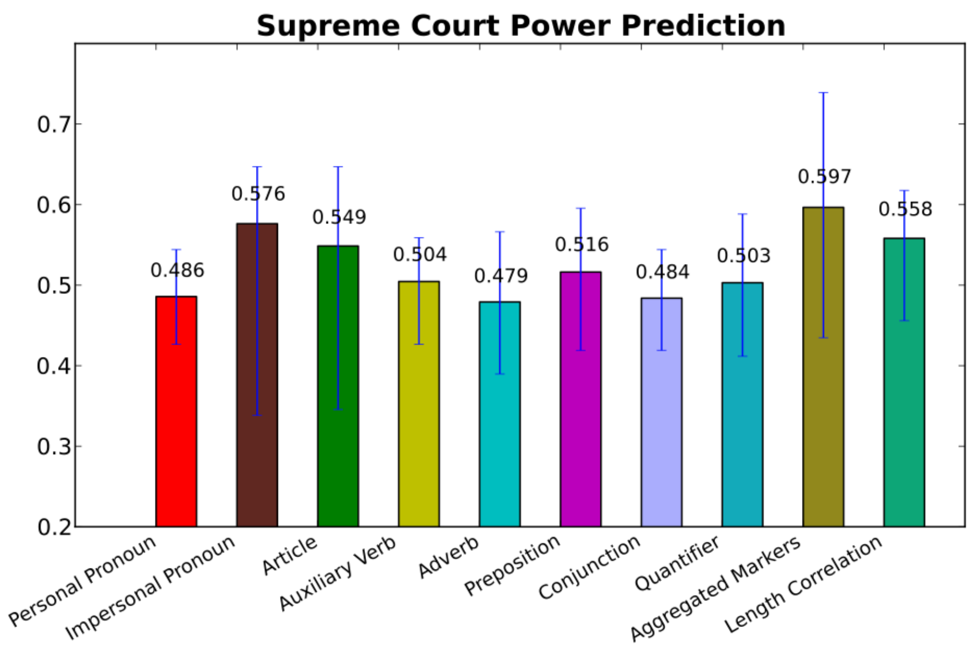


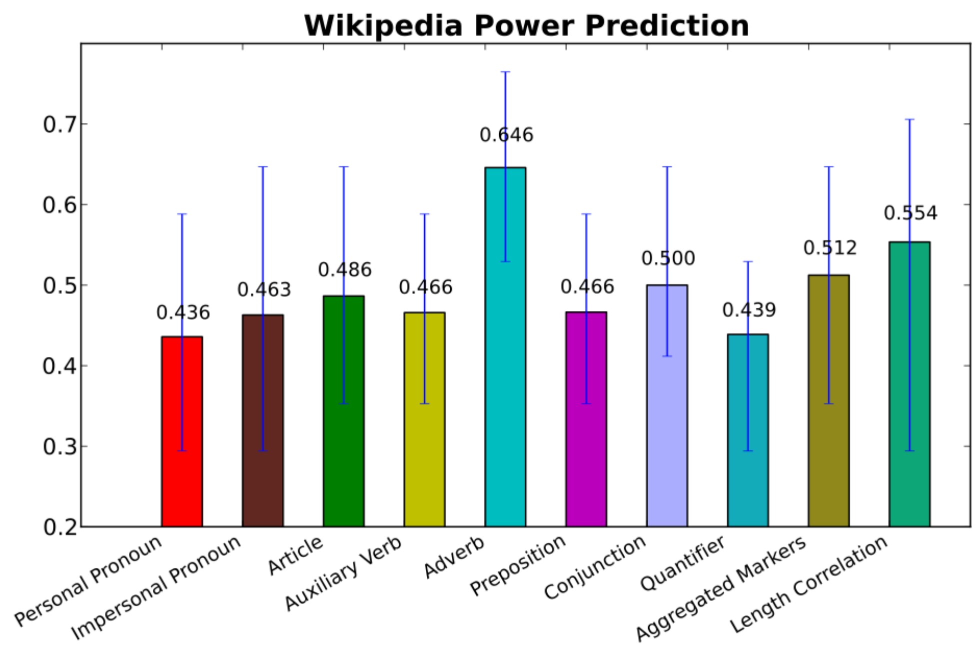


most “stylistic” coordination is “explained away” by length

Stylistic coordination and social status

- Can we use asymmetry in stylistic coordination to predict power relationship?
  - Justices vs. lawyers, admin vs. non-admins

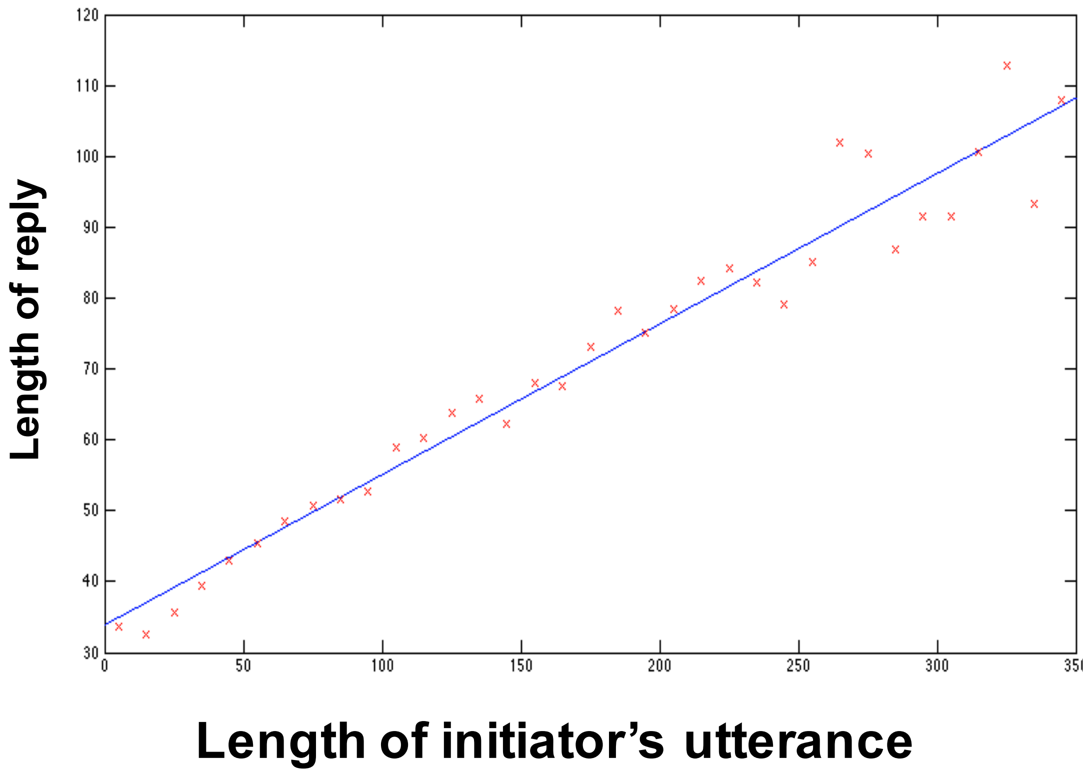




- Not really: observed asymmetry in stylistic coordination diminishes after conditioning on length

Length as a confounding factor

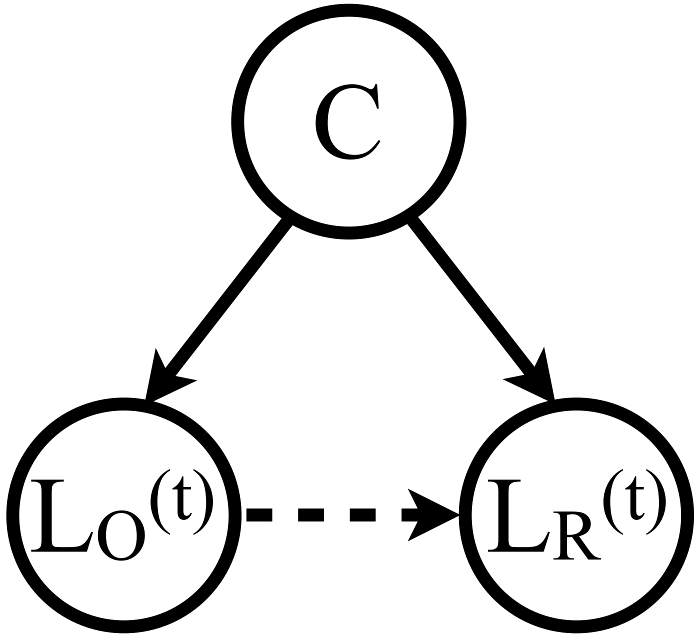
Wikipedia:



Longer utterances solicit longer response, producing spurious correlations in other features, e.g., # of occurrences of letter “r”

Understanding Length Coordination

- Bayesian Network for length coordination:



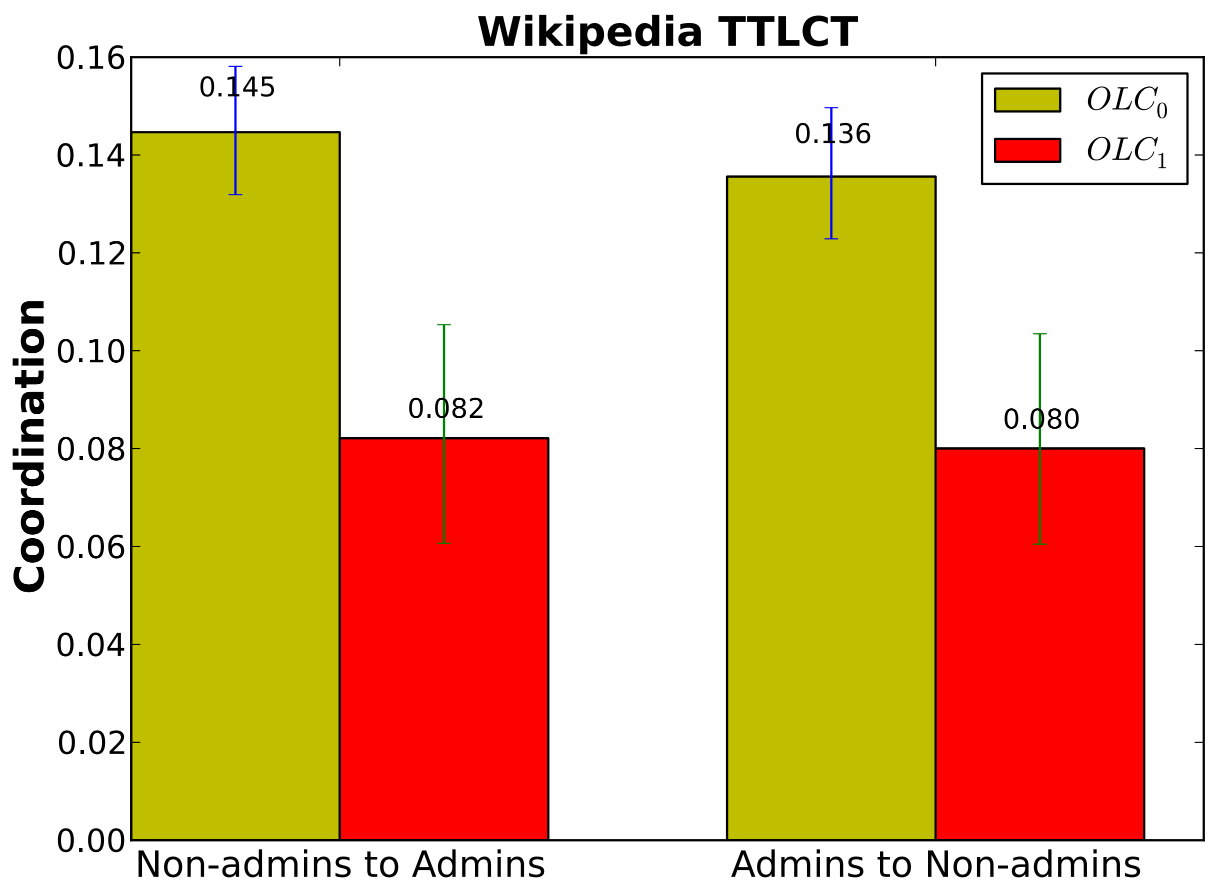
- Contextual factor: **C**
- Contextual influence: $\mathbf{C \rightarrow L_O}$ $\mathbf{C \rightarrow L_R}$
- Turn-by-turn length coordination: $\mathbf{L_O \rightarrow L_R}$

Turn-by-turn Length Coordination Test

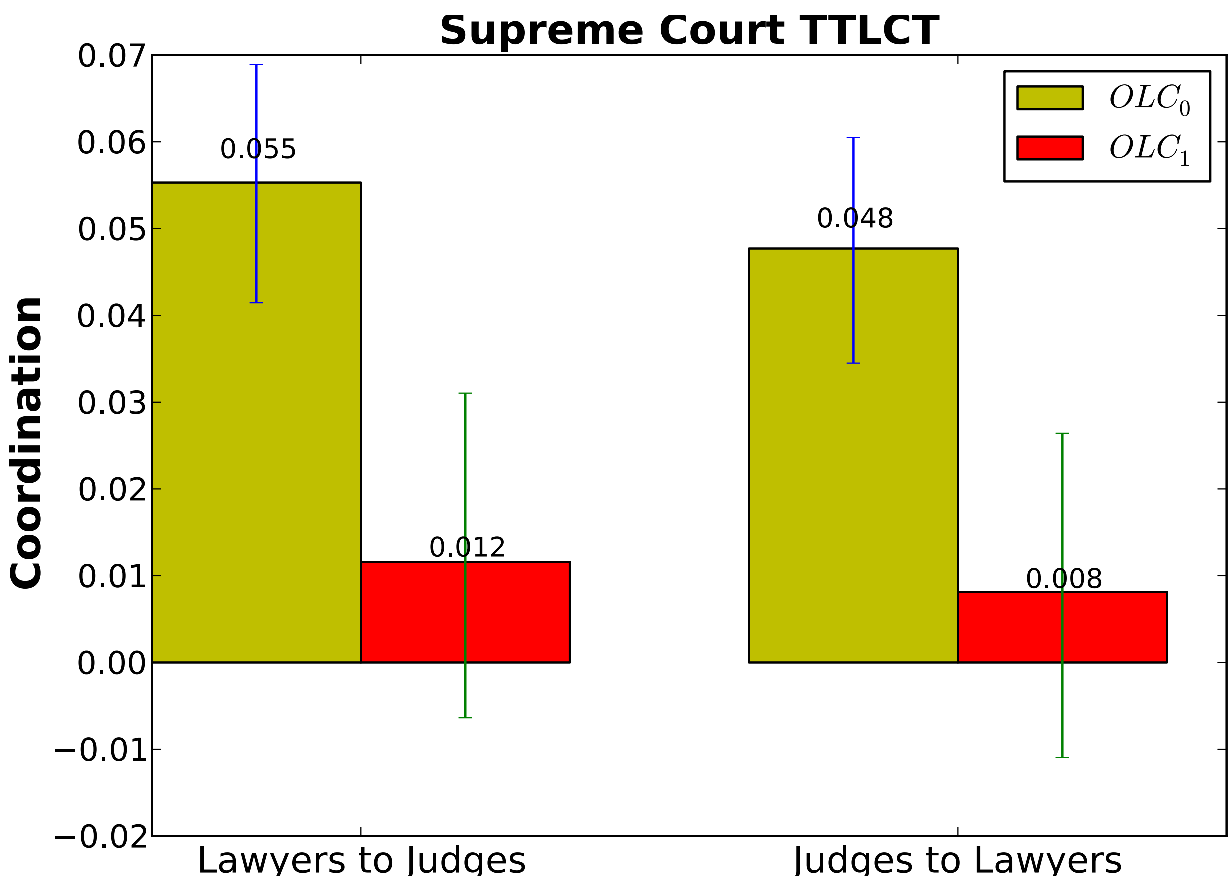
- A Conditional Monte Carlo Test
- *Overall Length Coordination*: OLC $=I(L_O:L_R)$
  - OLC_0: Original OLC
  - OLC_1: After shuffling utterances within each conversation
- Test: $OLC_0=OLC_1$?
  - If yes, then there is no turn-by-turn coordination

L_O	L_R	L_R
6	10	7
4	7	10
5	8	16
10	16	8

Turn-by-turn Length Coordination Test



Turn-by-turn Length Coordination Test



- Information Theory Basics
  - Entropy, MI, Discrete IT estimators
  - Entropy estimation demo
- Human behavior dynamics
  - Social networks
  - Stylistic coordination

Coffee Break (3:15-3:30)

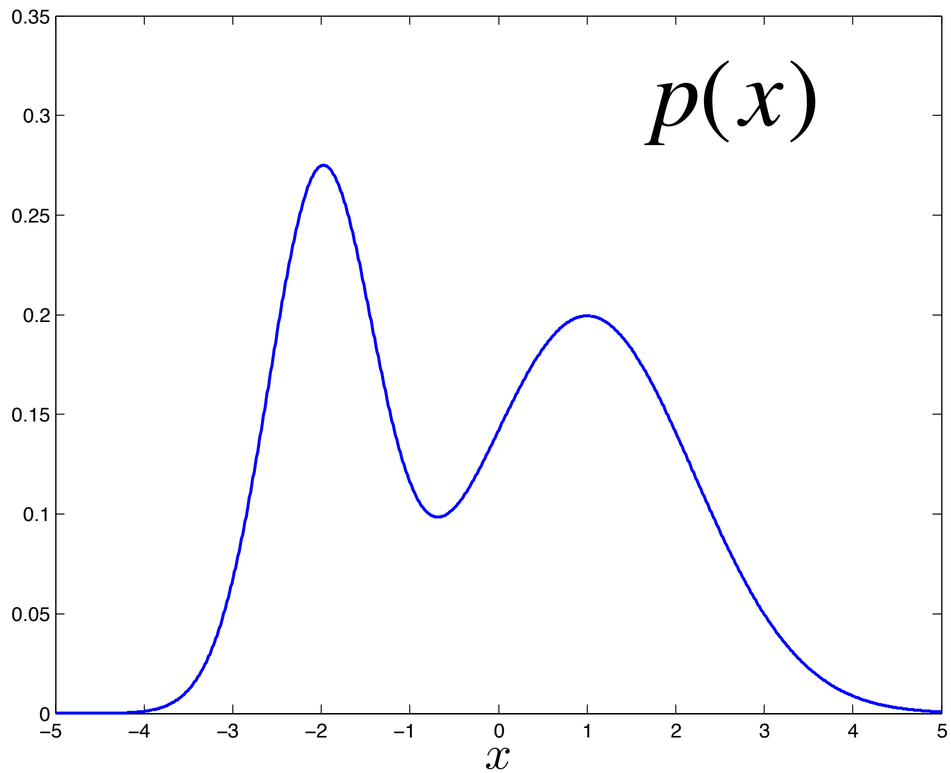
- **Non-parametric entropy estimation**
- Very high-dimensional information
  - How to handle it?
  - Applications: language, personality, behavior

Estimation of Entropic Measures from Data

Estimating Entropic Measures

$$H(X) = -\int dx p(x) \log p(x)$$

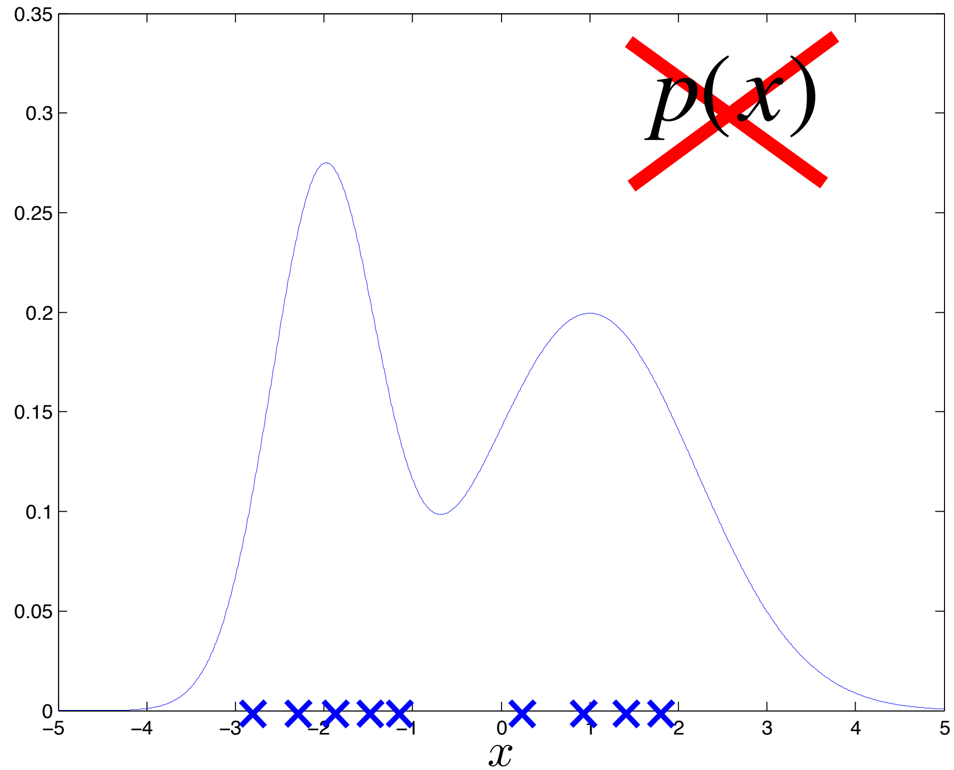
- Straightforward (kind of) if we know p(x)



Estimating Entropic Measures

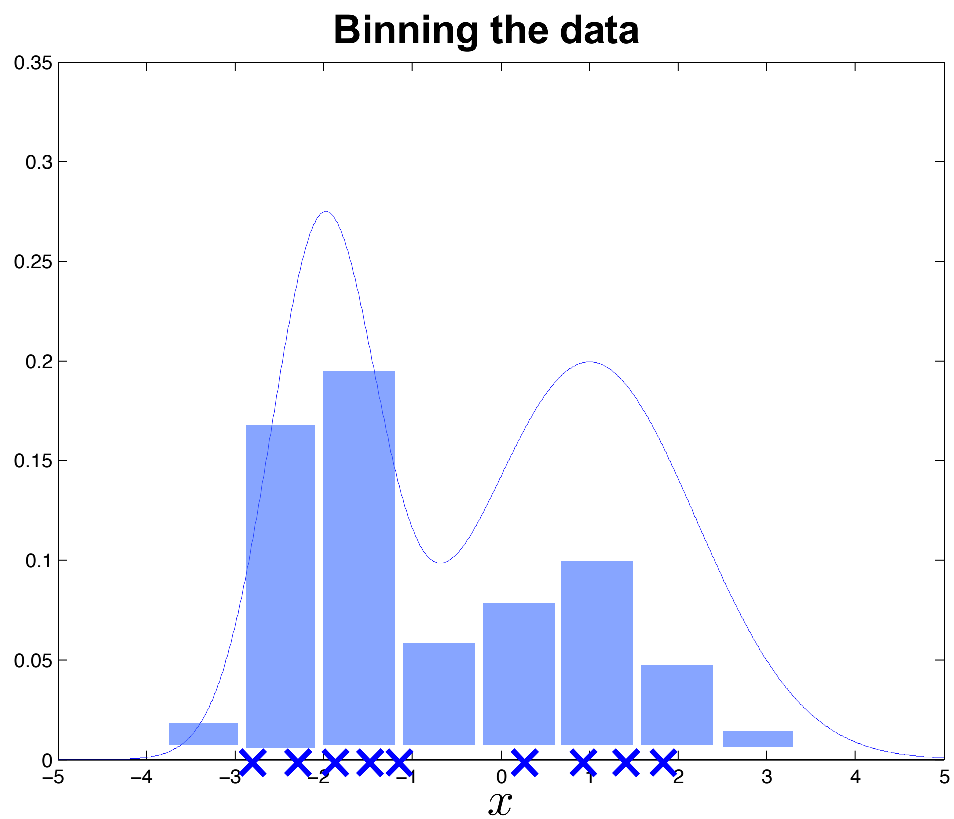
$$H(X) = -\int dx p(x) \log p(x)$$

- Usually we don't know p(x) (have samples $x_i \sim p(x)$)



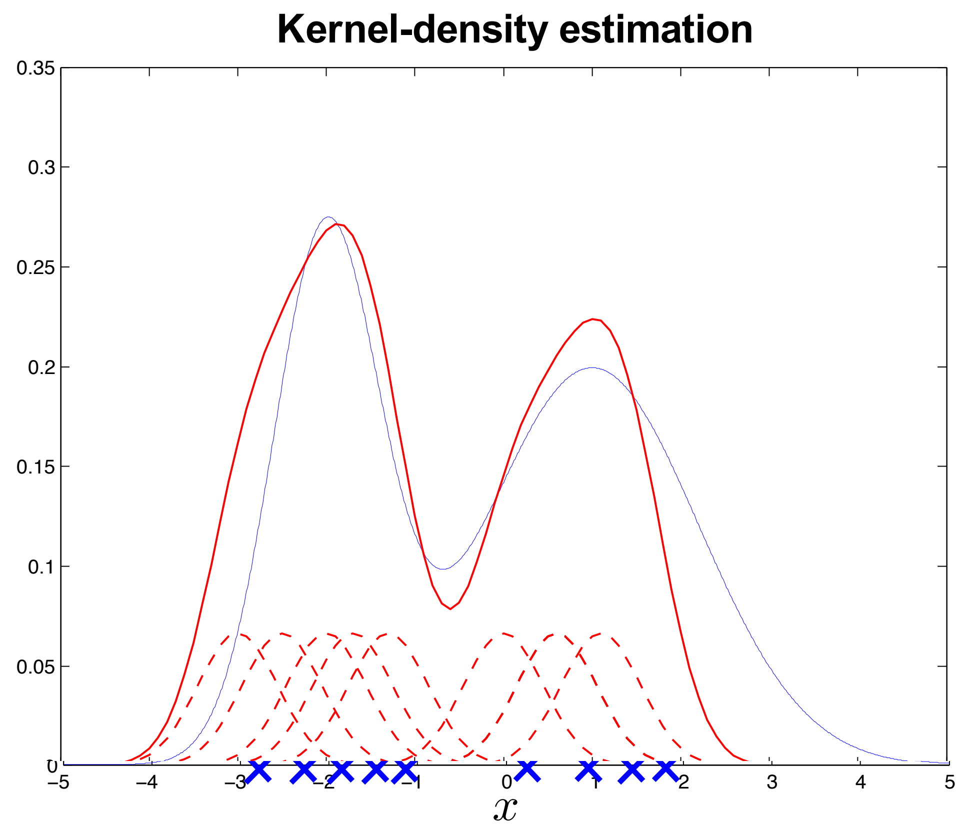
Plug-in Estimators

- Estimate p(x) and calculate the integral



Plug-in Estimators

- Estimate p(x) and calculate the integral



Does not work in high-dimensional, under-sampled settings

Binless Entropy Estimation

- One way to write entropy:

$$H(x) = \mathbb{E}_x[-\log p(x)]$$

- Given some samples $x_i \sim p(x)$,

$$\approx -\frac{1}{N}\sum_i \log p(x_i)$$

- We still don't know p(x)

- However, we need to estimate p(x) only at points x_i

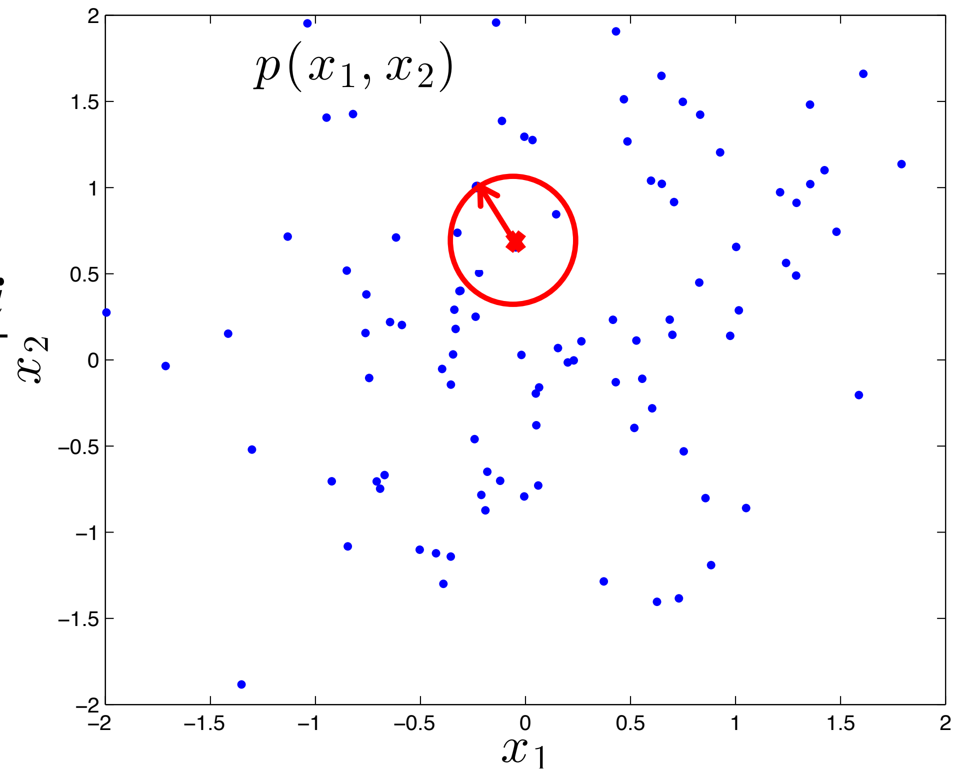
kNN Density Estimation for p(x)

- How to estimate the density p(x) at point $x^{(i)}$
  - Construct the k-nearest neighbor ball centered at $x^{(i)}$
  - **<u>Central Assumption</u>:**
    p(x) is uniform within the ball
  - Estimate

$$\hat{p}\left(\mathbf{x}^{(i)}\right) = \frac{\textit{probability mass of ball } i}{\textit{Volume of ball } i} = \frac{\textit{\% points in ball } i}{\textit{Volume of ball } i}$$

  - E.g. for d=2,k=4

$$\hat{p}_{k=4}\left(\mathbf{x}^{(i)}\right) = \frac{4/(N-1)}{\pi r_i^2}$$



$$\boxed{\hat{H}(\mathbf{x}) = -\frac{1}{N}\sum_{i=1}^{N}\log\hat{p}\left(\mathbf{x}^{(i)}\right) = \frac{2}{N}\sum_{i=1}^{N}\log r_i + \log(N-1) - \log k}$$

From Entropy to Mutual Information

- Mutual information is written as:

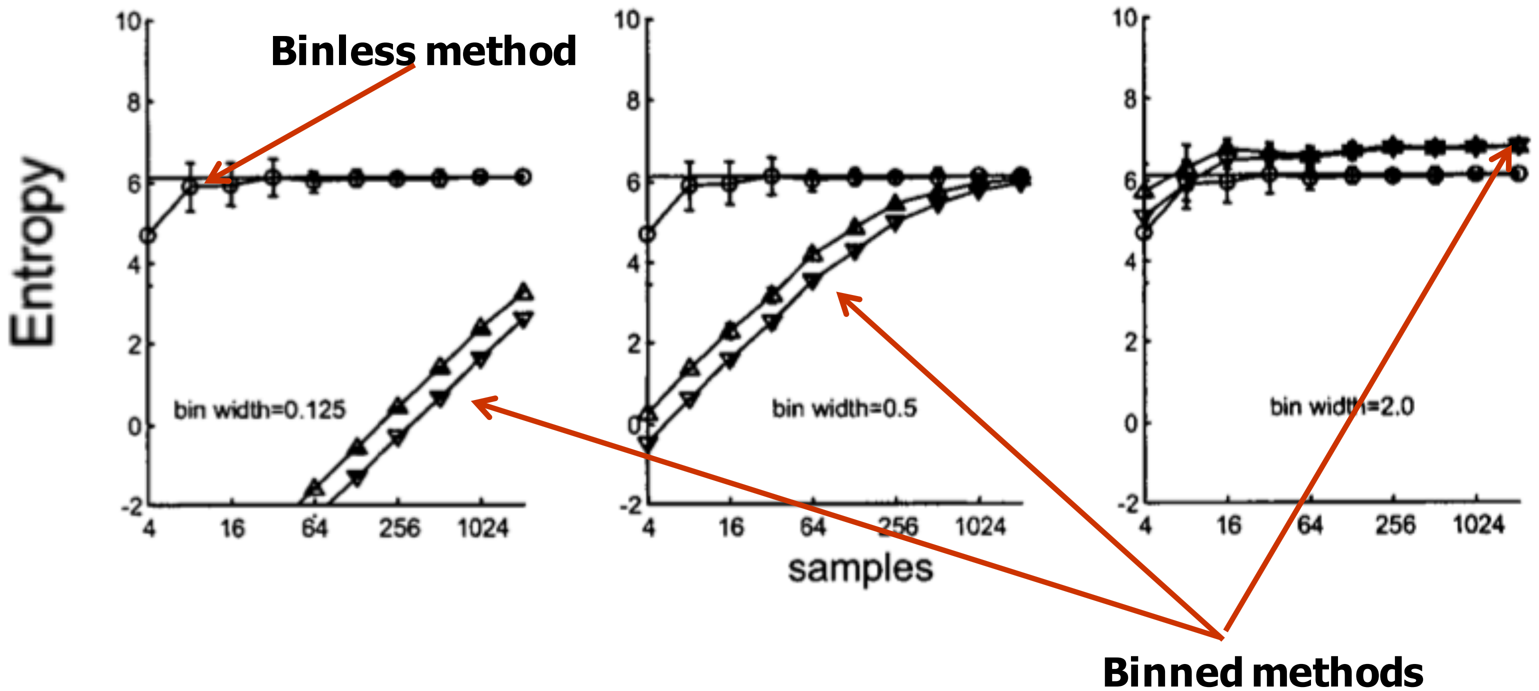
$$I(\mathbf{x}) = \sum_{i=1}^{d} H(x_i) - H(\mathbf{x})$$

- A simple MI estimator:

$$\hat{I}(\mathbf{x}) = \sum_{i=1}^{d} \widehat{H}(x_i) - \widehat{H}(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} \log \frac{\widehat{p}\left(\mathbf{x}^{(i)}\right)}{\widehat{p}\left(x_1^{(i)}\right)\widehat{p}\left(x_2^{(i)}\right)\dots\widehat{p}\left(x_d^{(i)}\right)}$$

Binless Entropy Estimation

Differential entropy for a Gaussian in 3 dimensions, as a function of N, the number of samples



From Victor 2002, "Binless strategies for estimation of information for neural data"

But for Topic Models?

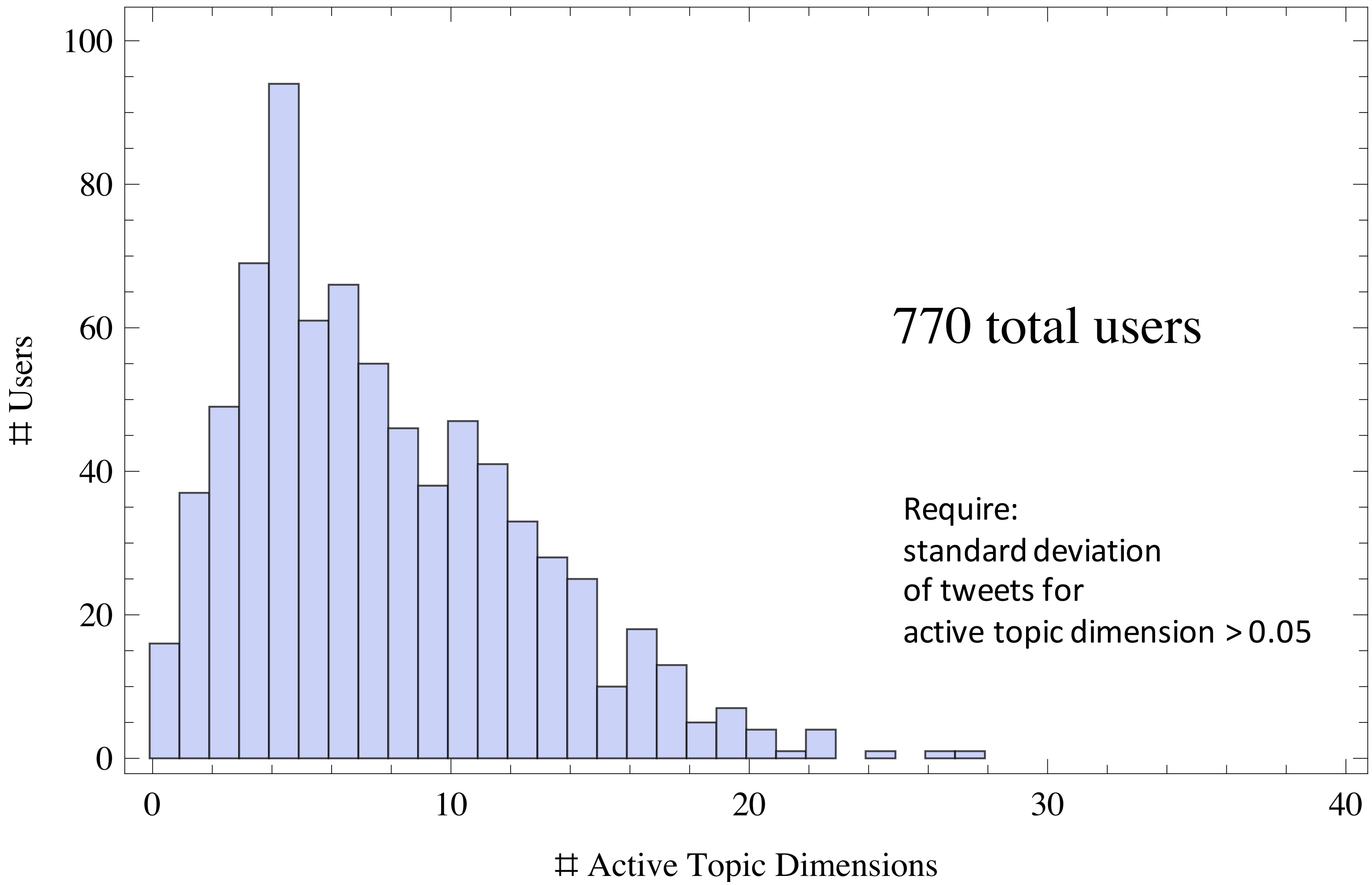
- Nice trick in a few dimensions, but if we pick a topic model with 125 topics,

$$X^{\mathrm{P}}, Y^{\mathrm{P}}, X^{\mathrm{F}} \in \mathbb{R}^{125}$$

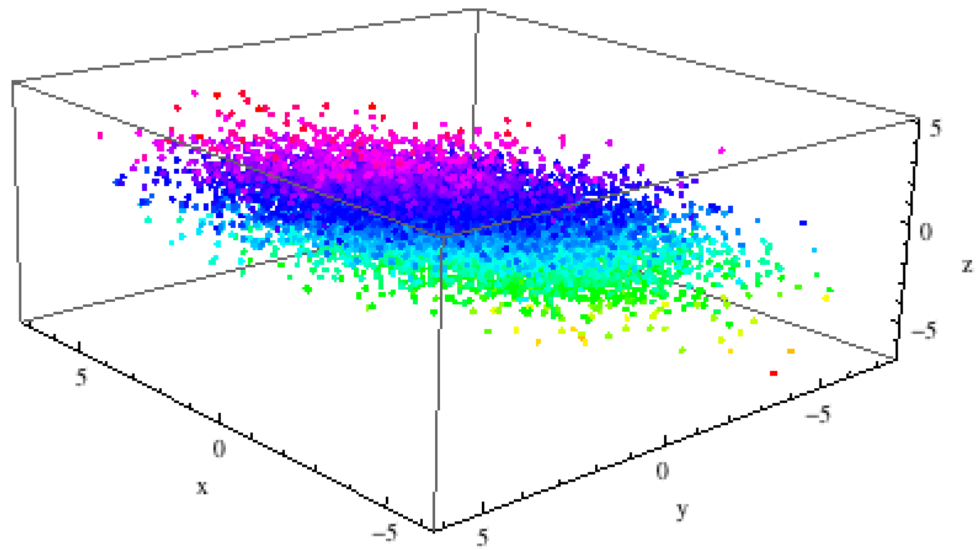
- Leads to a 375 dimensional space! We are estimating information transfer with as few as 100 samples!

- Ok, but is it REALLY 375 dimensional?
  - (answer: no! most people don't use most topics)

Number of active topics per user



Example

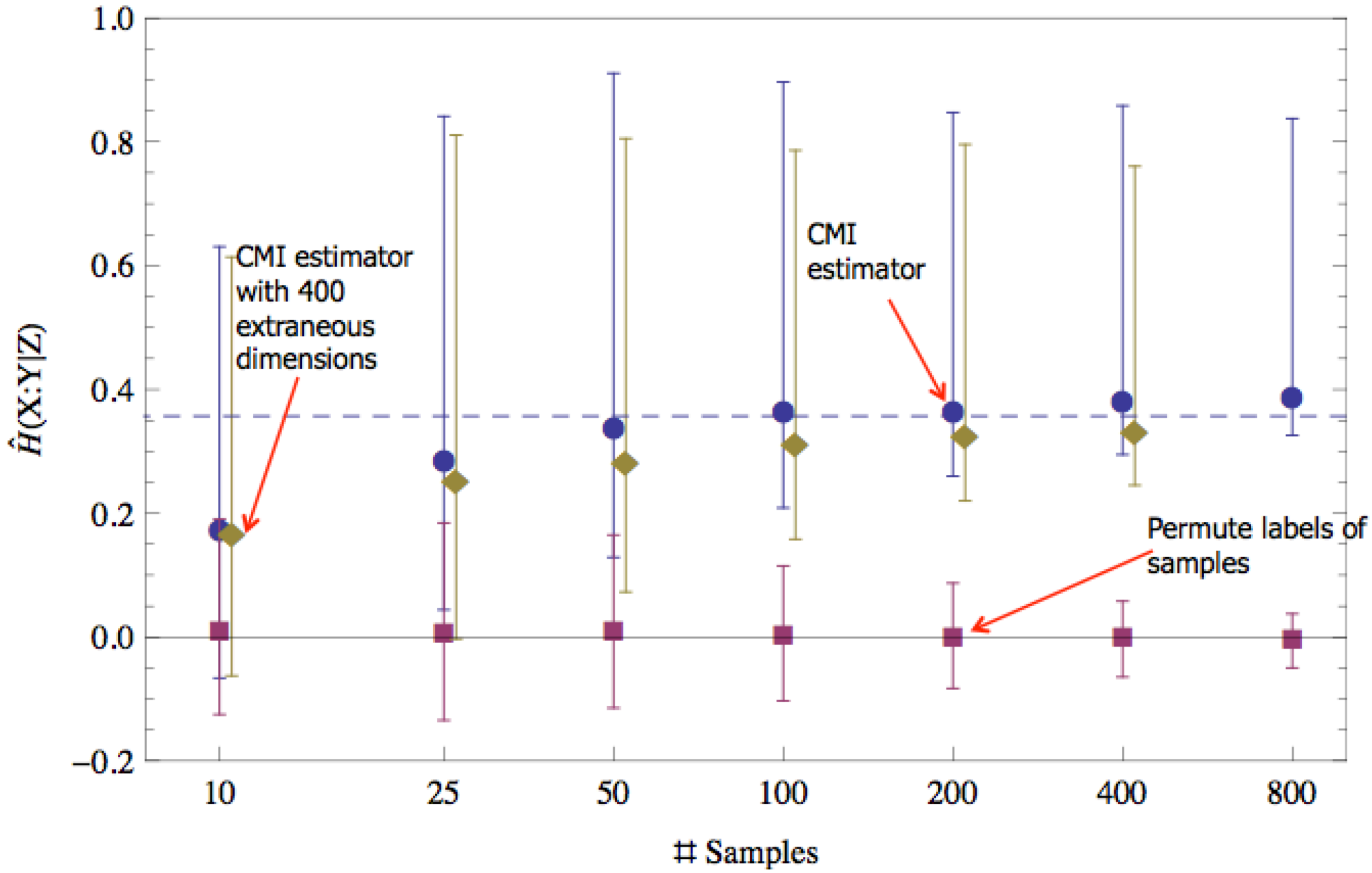


$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 4 & 3 & 1 \\ 3 & 4 & 1 \\ 1 & 1 & 2 \end{pmatrix}\right)$$

$$H(X:Y|Z) = 0.357$$

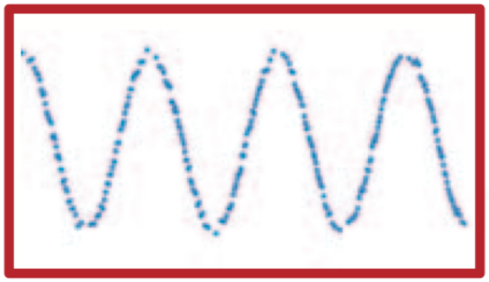
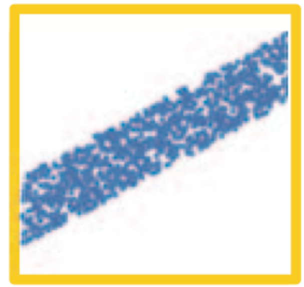
$$H(X:Y) = 0.413$$

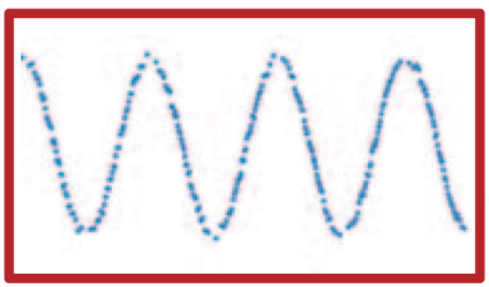
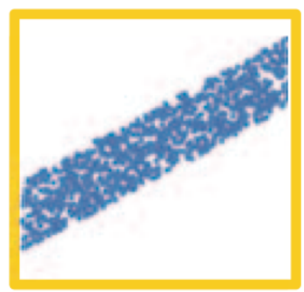
Convergence of estimators



Limitations of MI estimators

Reshef et al., “Detecting novel associations in large data sets.” Science, 2011

$\mathrm{MI}($  $) \neq \mathrm{MI}($  $) = 1.0$

$\widehat{\mathrm{MI}}($  $) \approx \widehat{\mathrm{MI}}($  $) \approx 1.0$

Mutual Information

PNAS

Equitability, mutual information, and the maximal information coefficient

Justin B. Kinney[1] and Gurinder S. Atwal

Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724

Edited* by David L. Donoho, Stanford University, Stanford, CA, and approved January 21, 2014 (received for review May 24, 2013)

How should one quantify the strength of association between two random variables without bias for relationships of a specific form? Despite its conceptual simplicity, this notion of statistical "equitability" has yet to receive a definitive mathematical formalization. Here we argue that equitability is properly formalized by a self-consistency condition closely related to Data Processing Inequality.

dependencies without bias for relationships of one type or another. And although it was proposed in the context of modeling communications systems, mutual information has been repeatedly shown to arise naturally in a variety of statistical problems (6–8).

The use of mutual information for quantifying associations in continuous data is unfortunately complicated by the fact that it

MI is just fine: one only needs more data points for accurate estimation

PNAS

Cleaning up the record on the maximal information coefficient and equitability

Although we appreciate Kinney and Atwal's interest in equitability and maximal information coefficient (MIC), we believe they misrepresent our work. We highlight a few of our main objections below.

Regarding our original paper (1), Kinney

instead that we look for approximations and solutions in restricted cases, an impossibility result about perfect equitability provides focus for further research, but does not mean that useful solutions are unattainable. Similarly, as others have noted

far will allow researchers in the area to most productively and collectively move forward.

David N. Reshef[a,b,1,2]***, Yakir A. Reshef***[b,1,2]***, Michael Mitzenmacher***[c,3]***, and Pardis C. Sabeti***[d,e,3]

PNAS

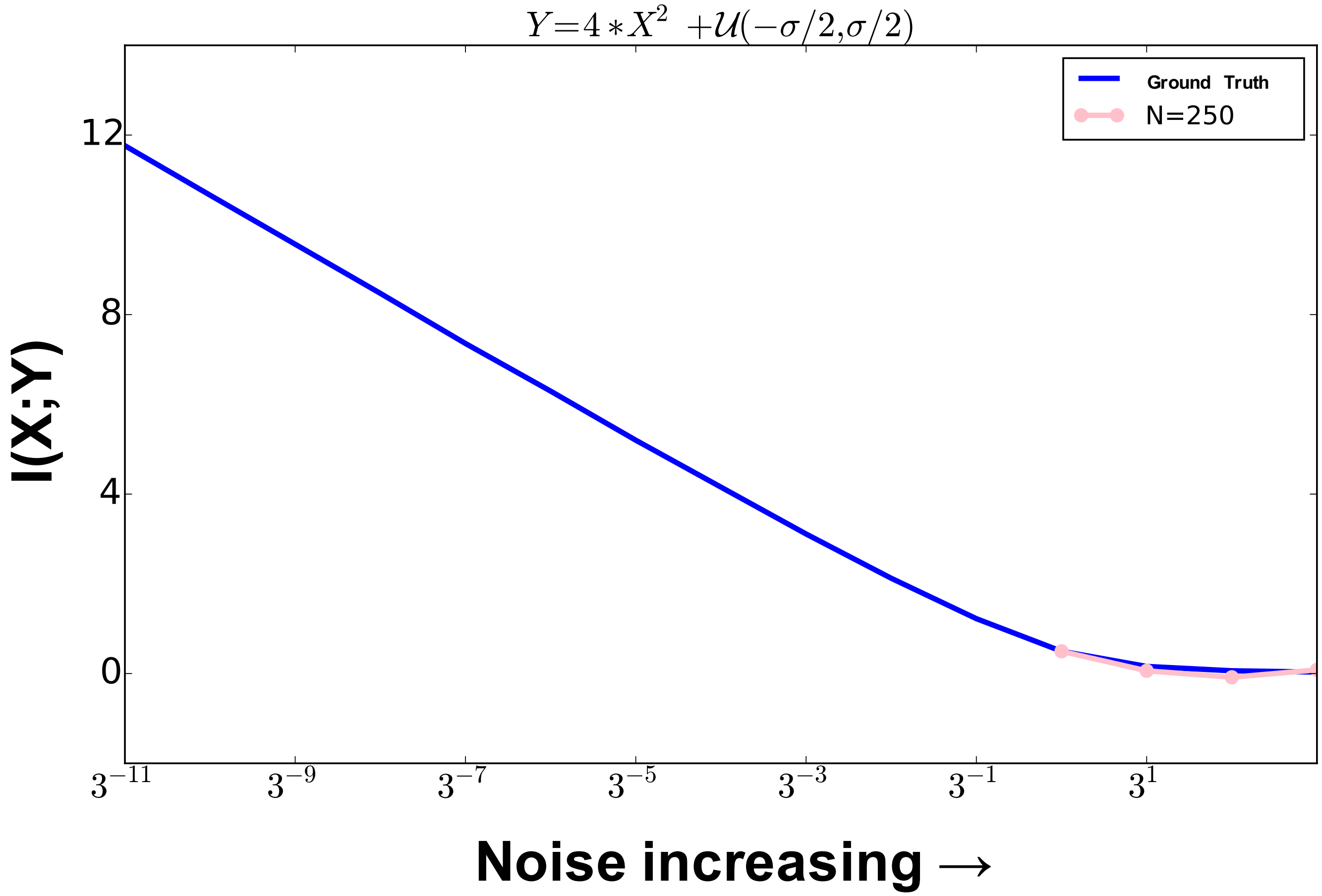
Reply to Reshef et al.: Falsifiability or bust

The term "equitability" was introduced by Reshef et al. in ref. 1 to describe measures of statistical dependence that "give similar scores to equally noisy relationships of different types." Their paper also introduced a new statistic, the "maximal information coefficient" (MIC), that was said to satisfy this equitability criterion. There has since been

the claimed equitability of MIC was only intended to describe a qualitative tendency that they observed when analyzing some data that they themselves simulated. We find this objection of theirs troubling, as it implies that the central claim of ref. 1—that MIC is equitable—was never meant to be falsifiable.

mately satisfy R^2-equitability better than do certain estimates of mutual information. The relevance of these select simulations is unclear. As proven in our paper, neither MIC nor mutual information satisfies R^2-equitability in any mathematical sense. The question of whether estimates of these quantities are approximately R^2-equitable is therefore nei-

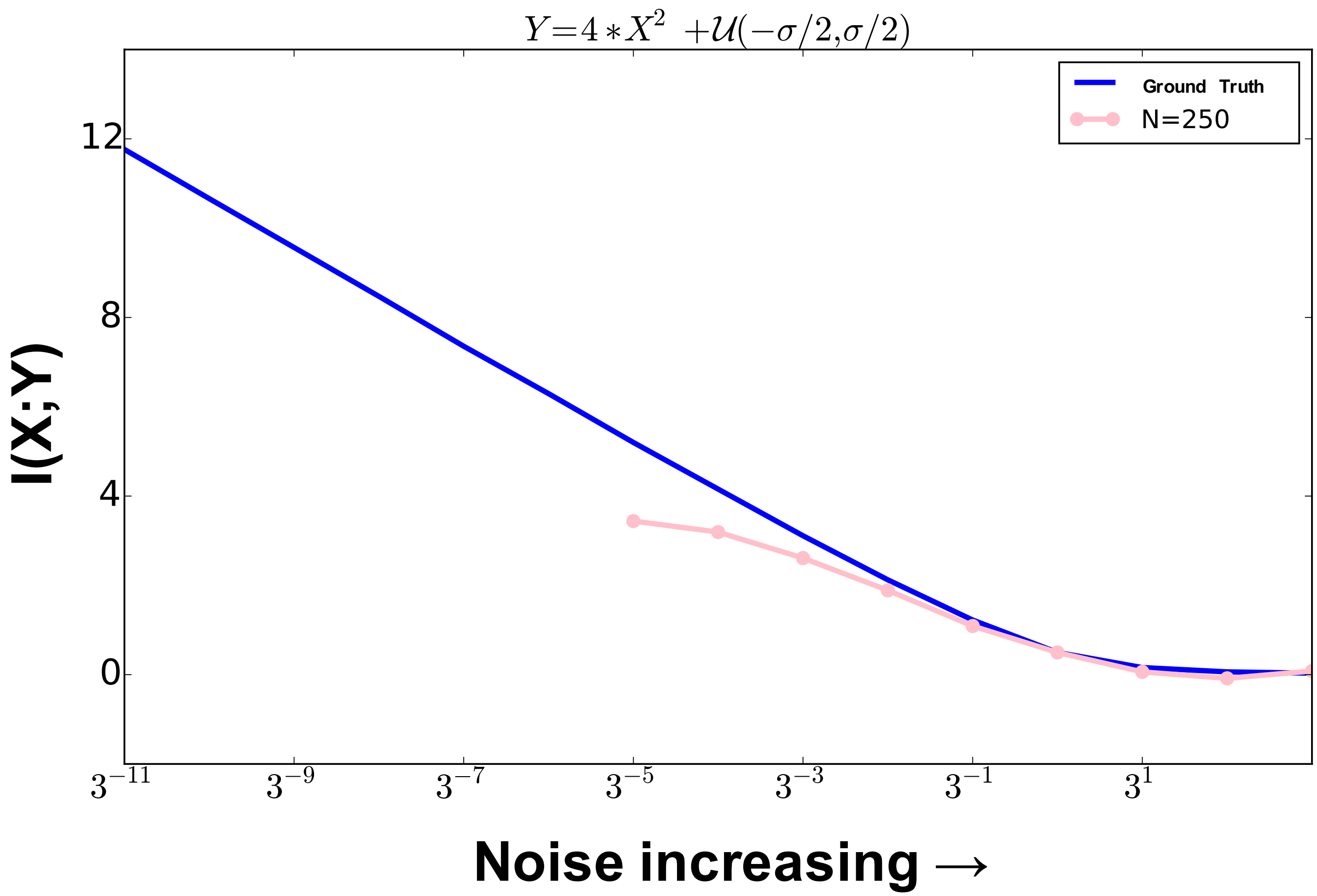
Mutual Information as a Function of Noise



Kraskov, Stögbauer, & Grassberger, Physical Review E, 2004

$$\boxed{\hat{I}_{KSG,k}(\mathbf{x}) = (d-1)\psi(N) + \psi(k) - (d-1)/k - \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{d}\psi\left(n_{x_j}(i)\right)}$$

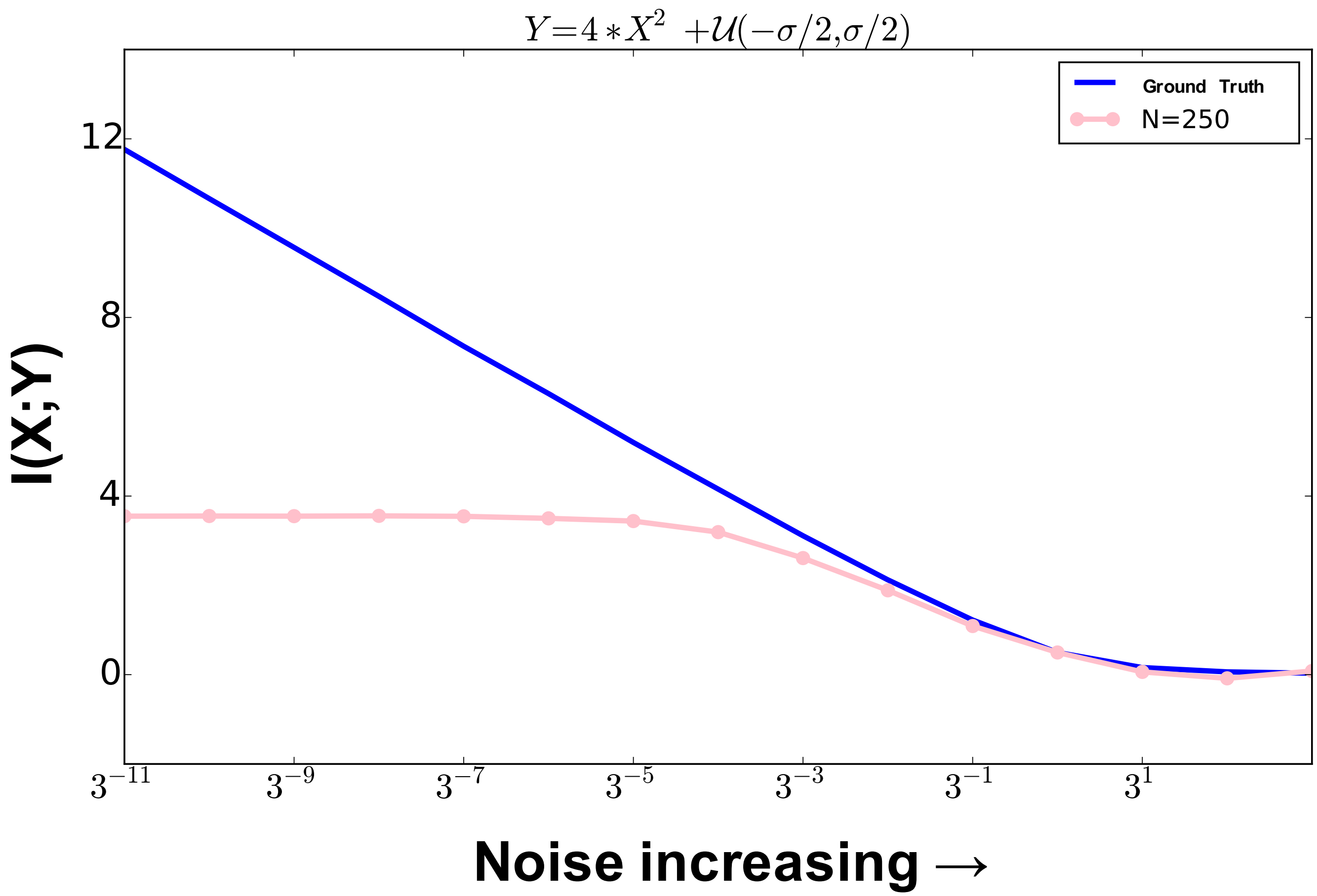
Mutual Information as a Function of Noise



Kraskov, Stögbauer, & Grassberger, Physical Review E, 2004

$$\hat{I}_{KSG,k}(\mathbf{x}) = (d-1)\psi(N) + \psi(k) - (d-1)/k - \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{d}\psi\left(n_{x_j}(i)\right)$$

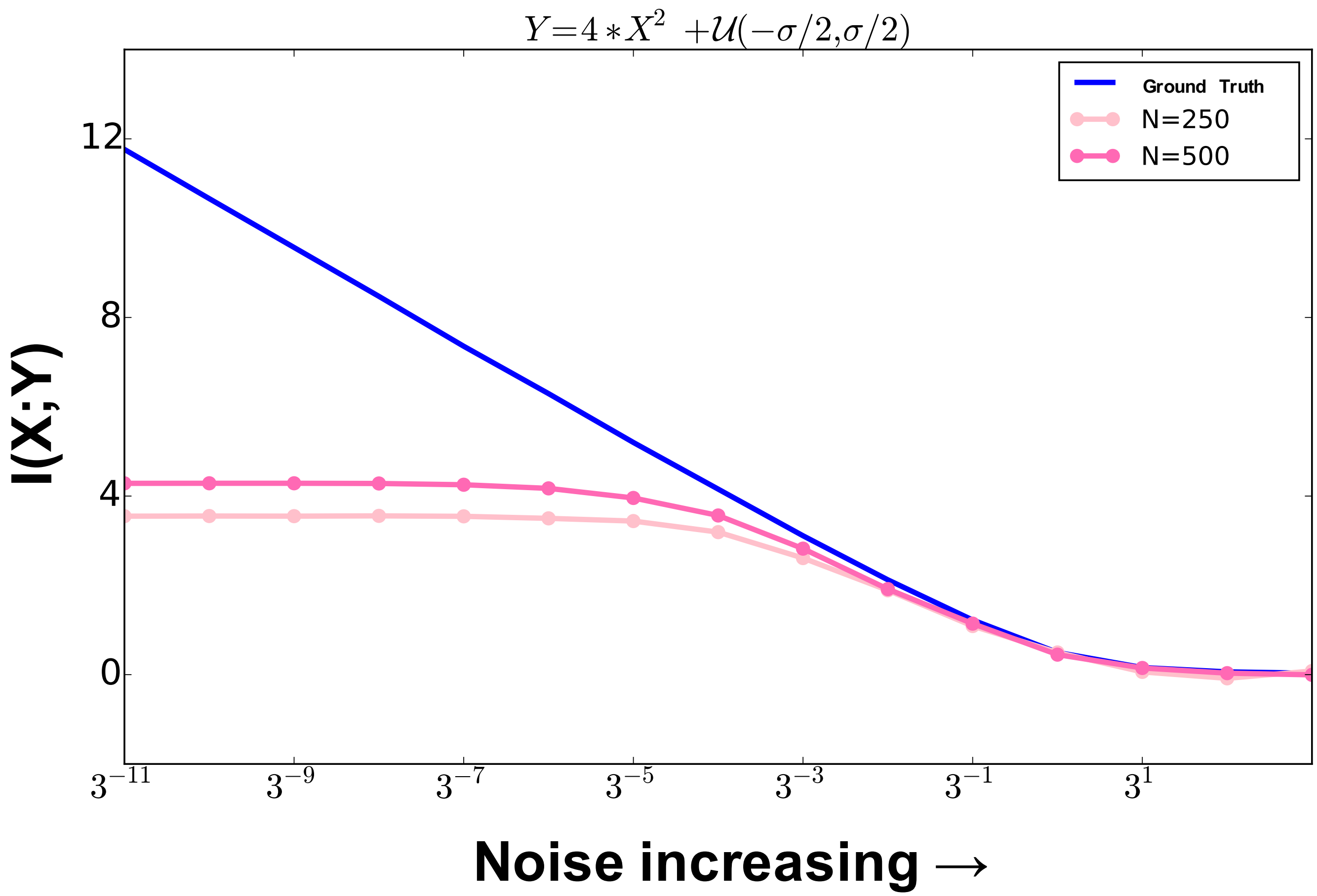
Mutual Information as a Function of Noise



Kraskov, Stögbauer, & Grassberger, Physical Review E, 2004

$$\hat{I}_{KSG,k}(\mathbf{x}) = (d-1)\psi(N) + \psi(k) - (d-1)/k - \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{d}\psi\left(n_{x_j}(i)\right)$$

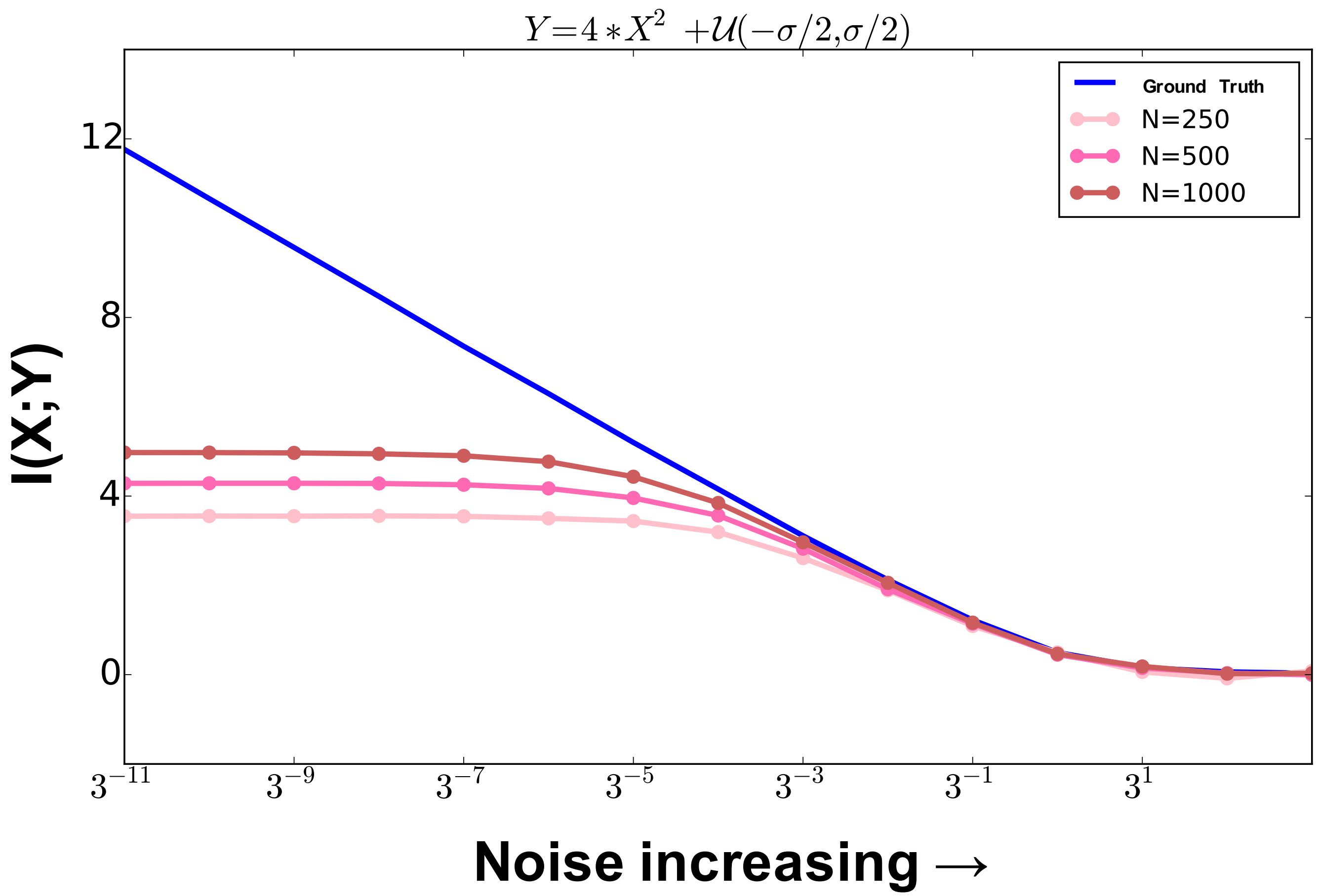
Mutual Information as a Function of Noise



Kraskov, Stögbauer, & Grassberger, Physical Review E, 2004

$$\hat{I}_{KSG,k}(\mathbf{x}) = (d-1)\psi(N) + \psi(k) - (d-1)/k - \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{d}\psi\left(n_{x_j}(i)\right)$$

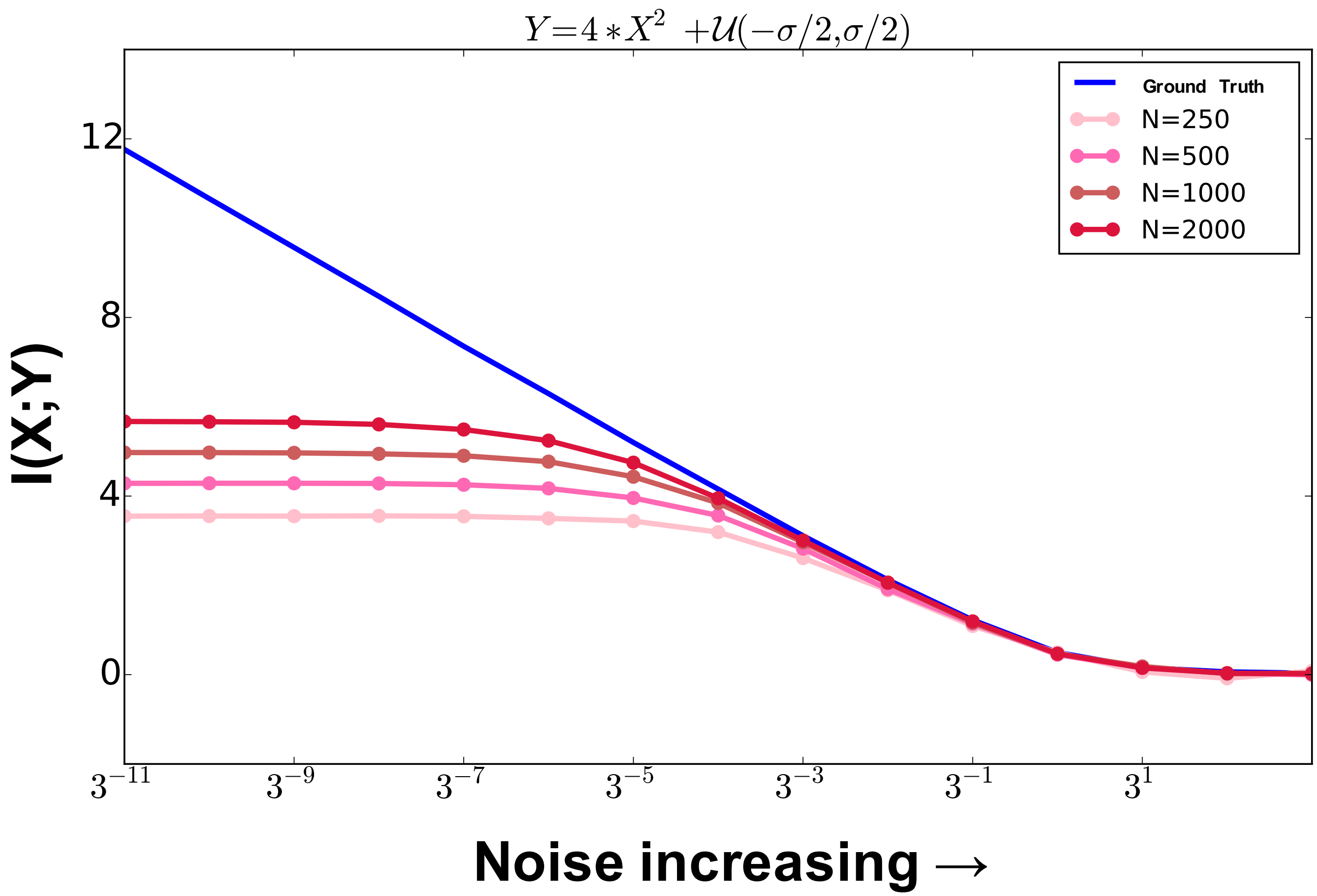
Mutual Information as a Function of Noise



Kraskov, Stögbauer, & Grassberger, Physical Review E, 2004

$$\hat{I}_{KSG,k}(\mathbf{x}) = (d-1)\psi(N) + \psi(k) - (d-1)/k - \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{d}\psi\left(n_{x_j}(i)\right)$$

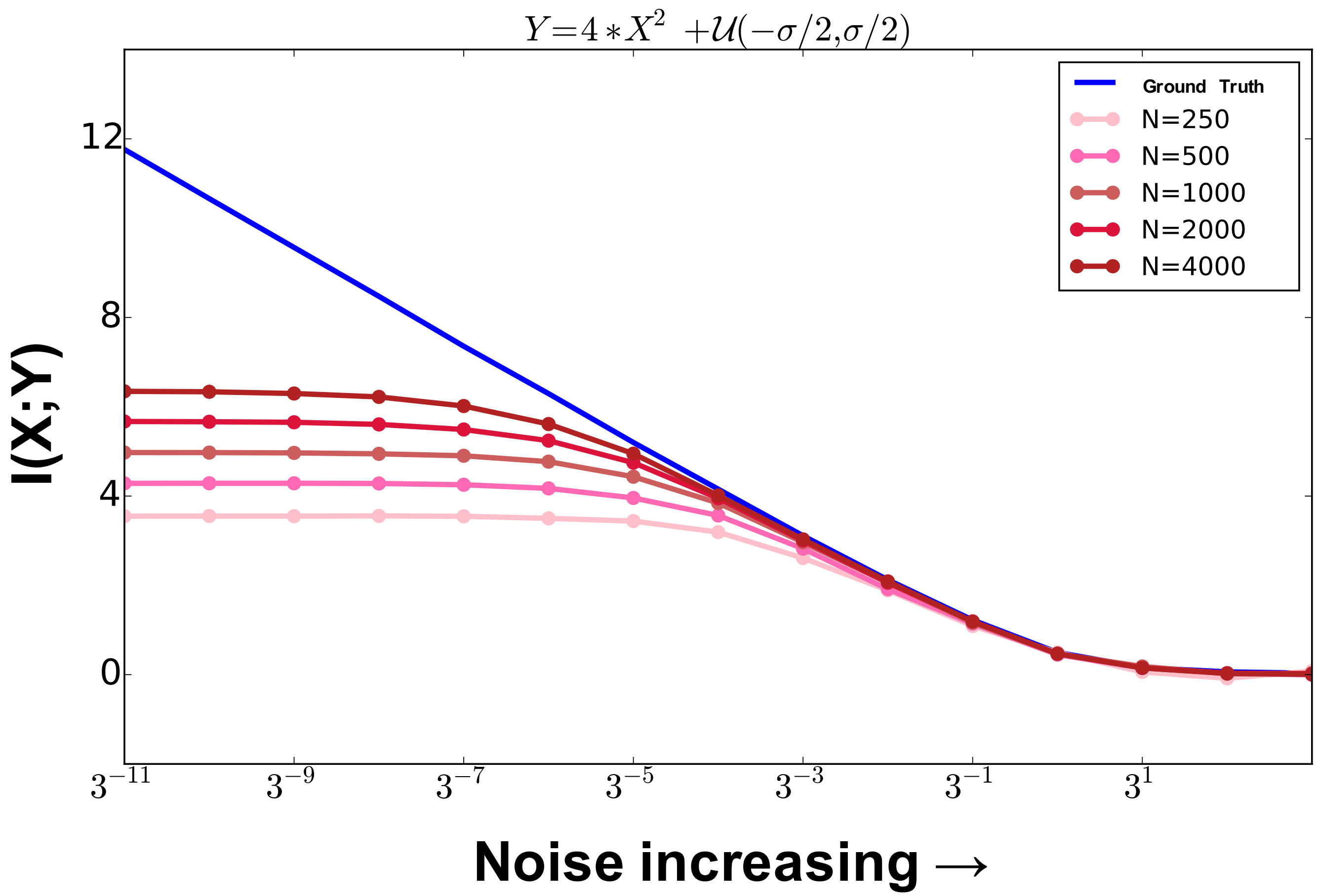
Mutual Information as a Function of Noise



Kraskov, Stögbauer, & Grassberger, Physical Review E, 2004

$$\boxed{\hat{I}_{KSG,k}(\mathbf{x}) = (d-1)\psi(N) + \psi(k) - (d-1)/k - \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{d}\psi\left(n_{x_j}(i)\right)}$$

Mutual Information as a Function of Noise



Kraskov, Stögbauer, & Grassberger, Physical Review E, 2004

$$\boxed{\hat{I}_{KSG,k}(\mathbf{x}) = (d-1)\psi(N) + \psi(k) - (d-1)/k - \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{d}\psi\left(n_{x_j}(i)\right)}$$

kNN Estimator Limitations

Theorem

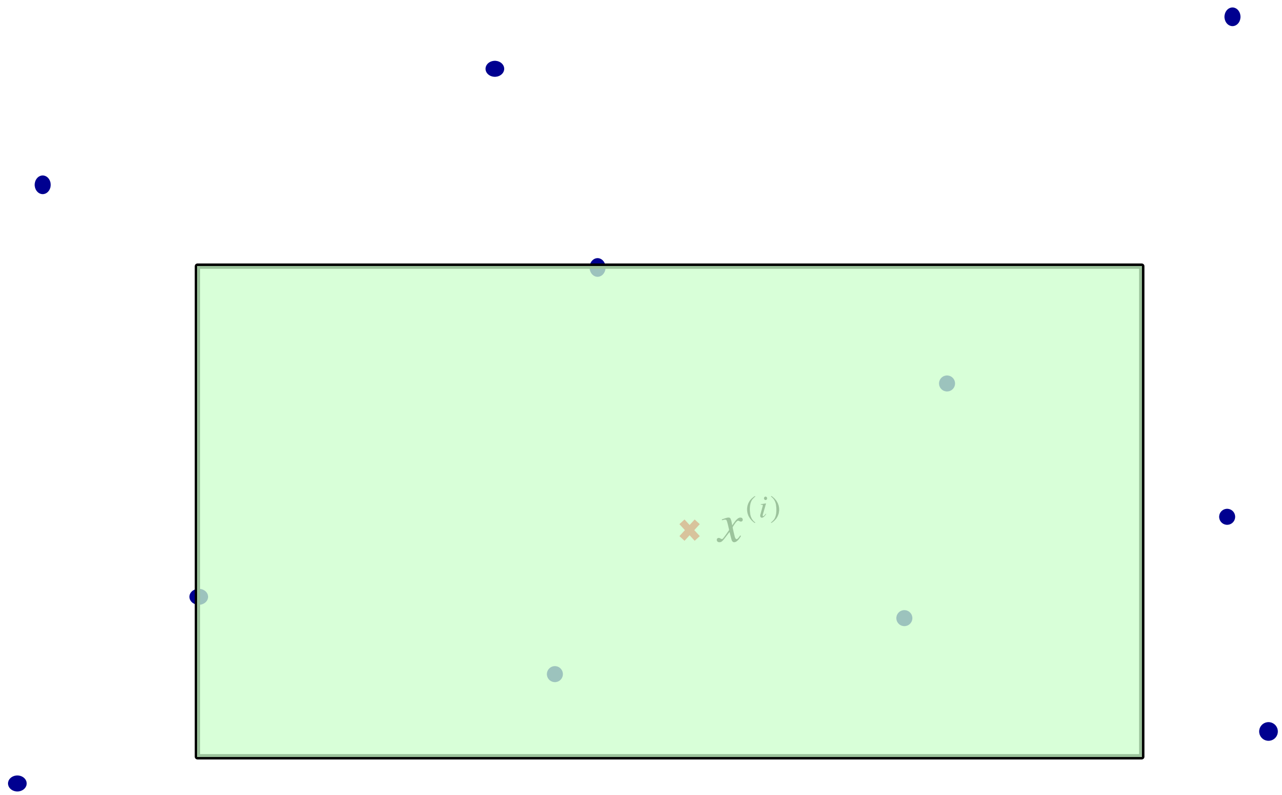
For a certain class of k-NN estimators, estimating mutual information within ε of its true value, $\left|\hat{I}(\mathbf{x}) - I(\mathbf{x})\right| \le \varepsilon$, requires that the number of samples, N, is at least:

$$N \ge C \exp\left(\frac{I(\mathbf{x}) - \varepsilon}{d-1}\right) + 1$$

Strong relationships require exponentially many samples to measure

kNN Estimator Limitations

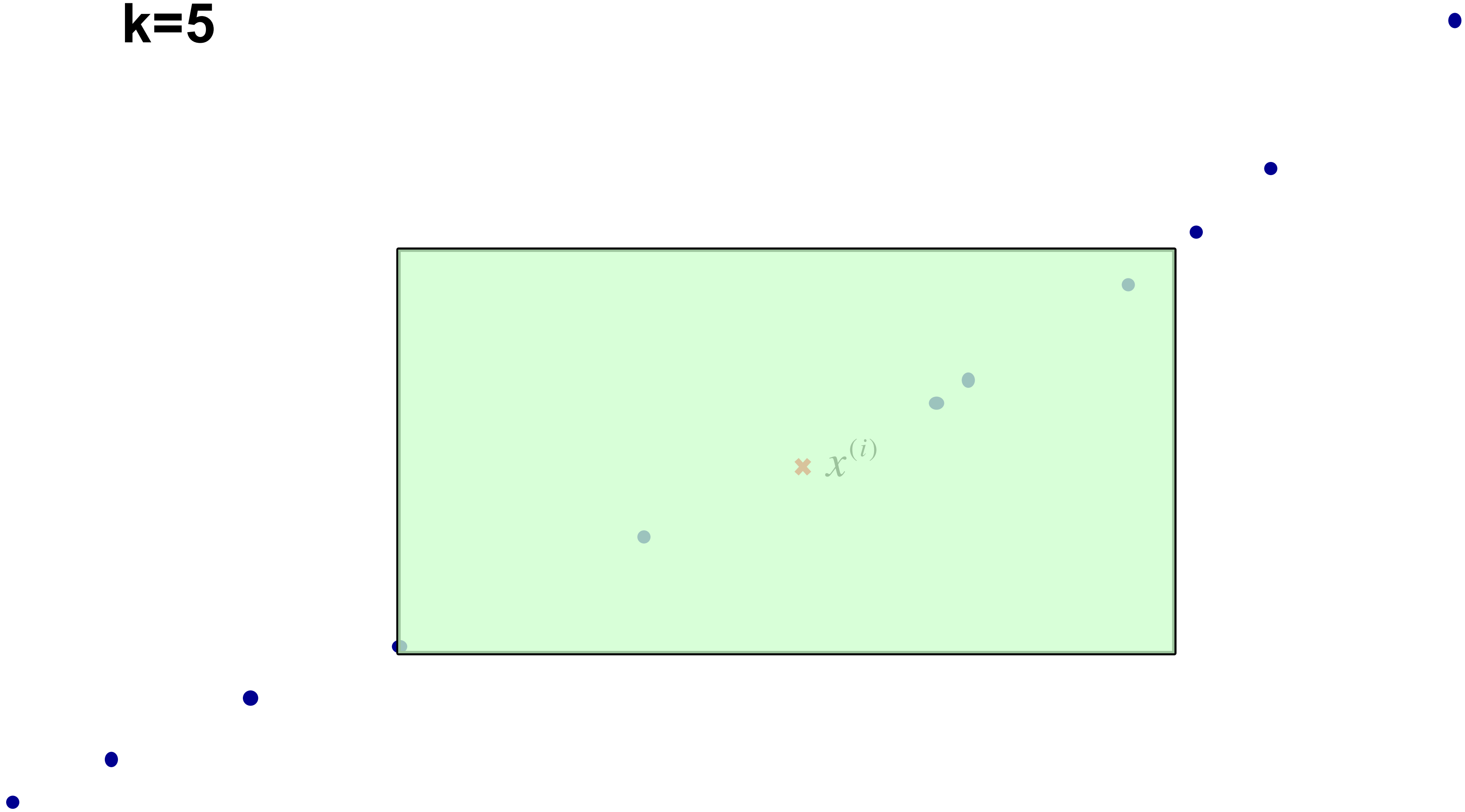
k=5



Works well for weakly correlated distributions

kNN Estimator Limitations

k=5

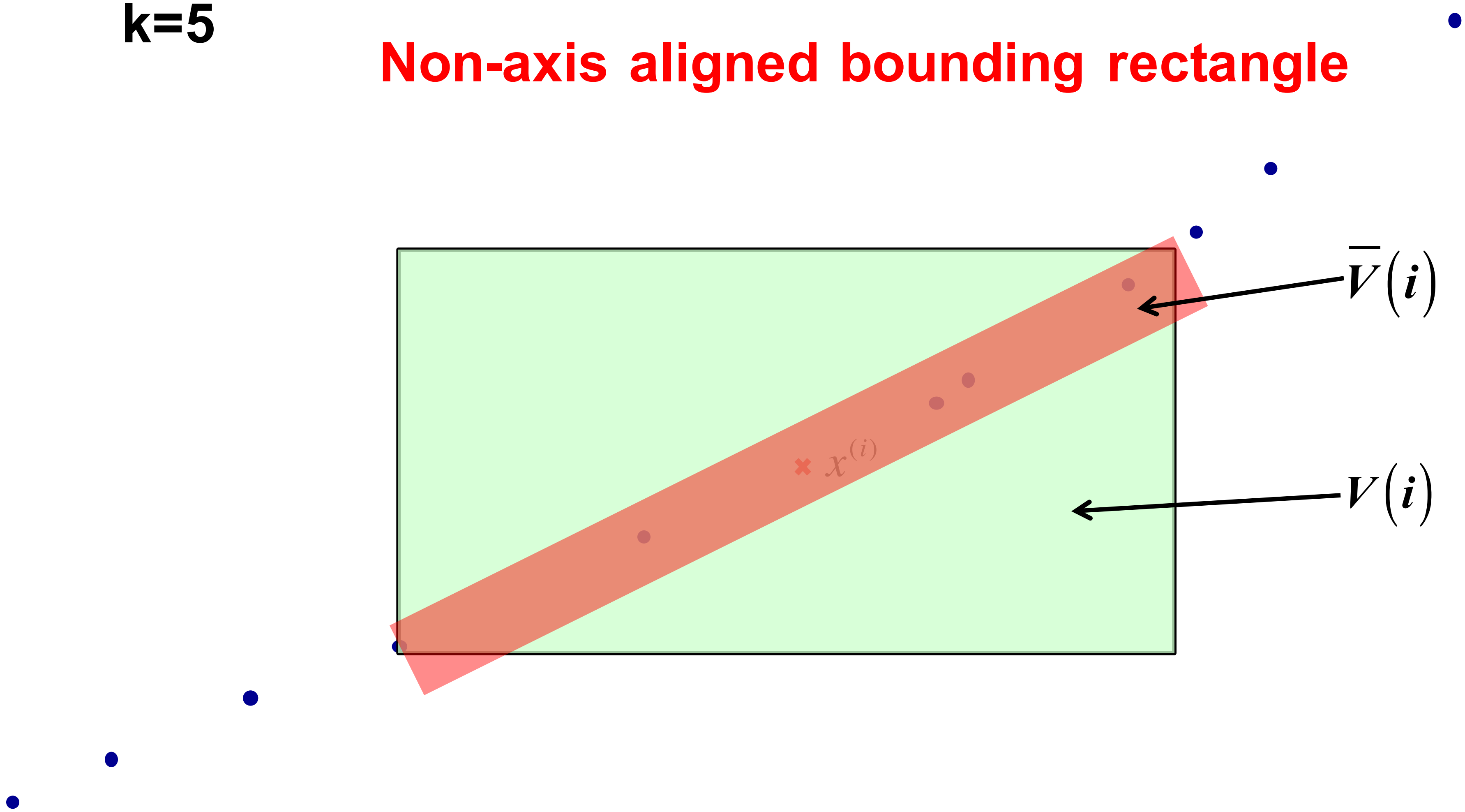


Works bad for strongly correlated distributions

Put a lot more probability mass out of the support

Relax Local Uniformity Condition

k=5



$$\hat{I}_{LNC}(\mathbf{x}) = \hat{I}(\mathbf{x}) - \frac{1}{N}\sum_{i=1}^{N}\log\frac{\bar{V}(i)}{V(i)}$$

Local Non-Uniform Correction Algorithm

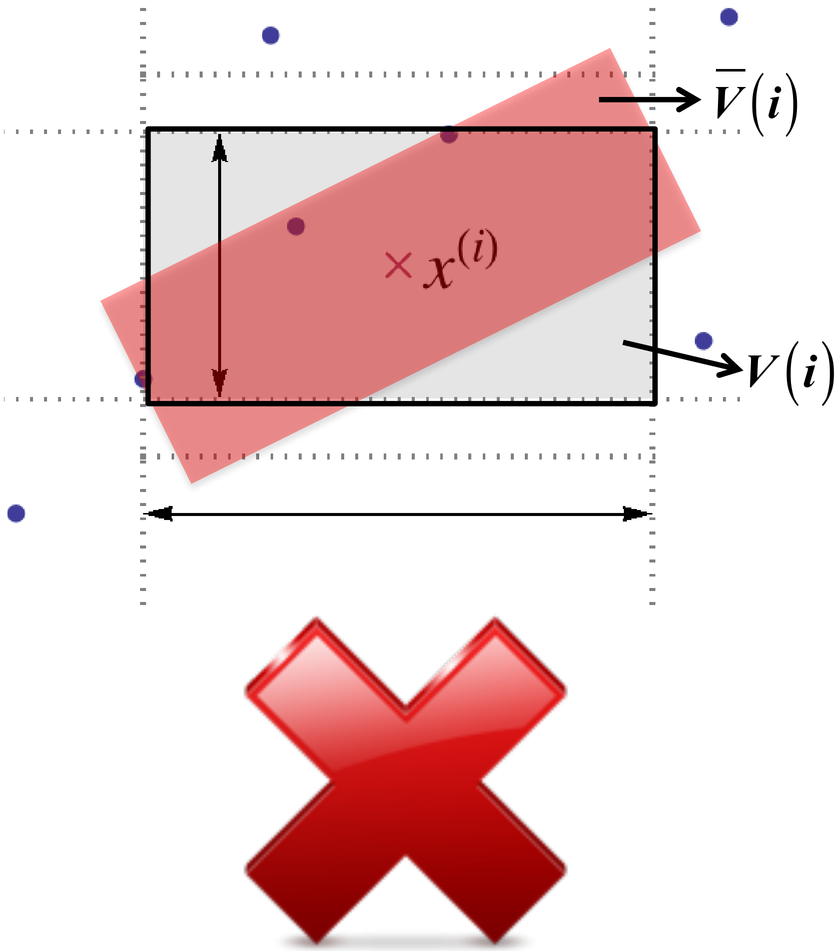
Algorithm 1 **Mutual Information Estimation with Local Nonuniform Correction**

$correction = 0$
for each point $\mathbf{x}^{(i)}$ **do**
  Find k nearest neighbors of $\mathbf{x}^{(i)}$
  Calculate volume of kNN rectangle $V(i)$
  Apply PCA on k neighbors, obtain volume $\overline{V}(i)$
  if $\overline{V}(i)/V(i) < \alpha_{k,d}$ **then**
    $correction = correction + \log \frac{\overline{V}(i)}{V(i)}$
  end if
end for
$\hat{I}_{LNC}(\mathbf{x}) = \hat{I}(\mathbf{x}) - \frac{1}{N} * correction$

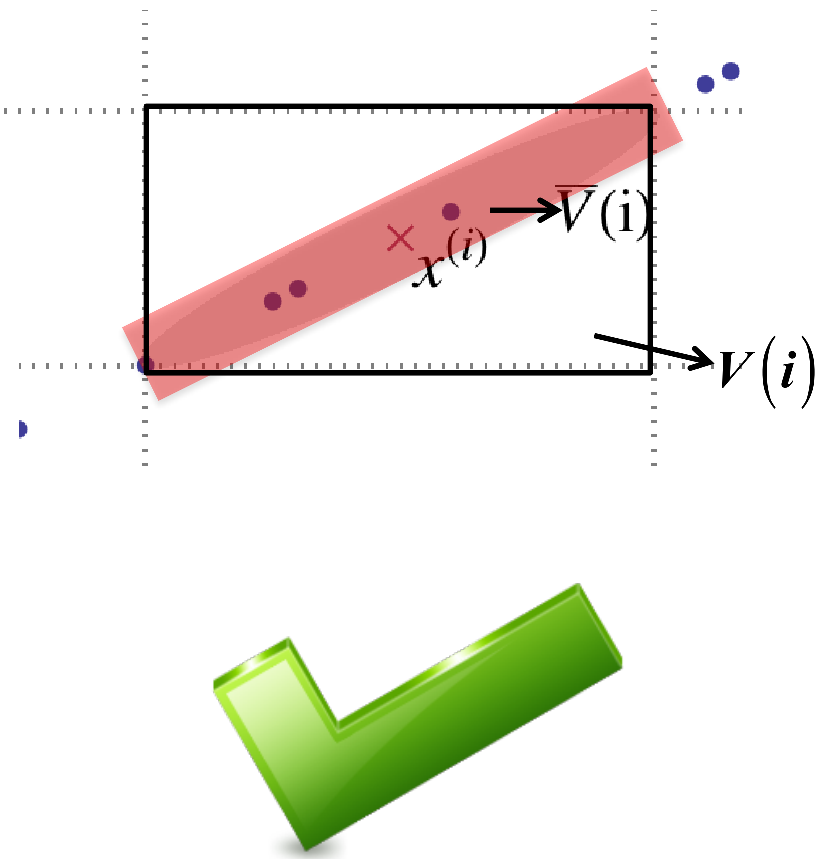
Non-Uniformity
Checking

Test for Local Non-Uniformity

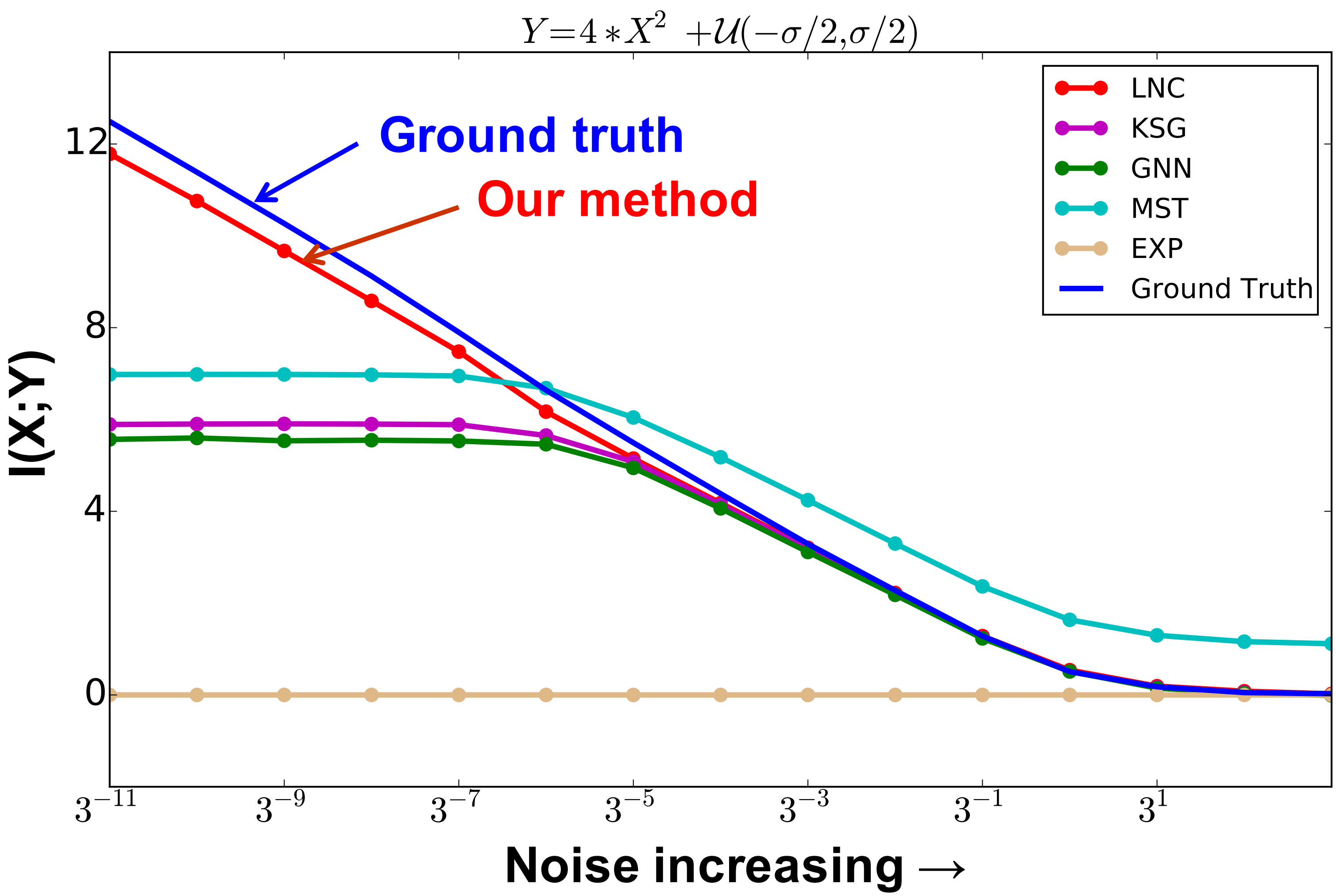
$\overline{V}(\boldsymbol{i})/V(\boldsymbol{i}) \geq \alpha_{k,d}$



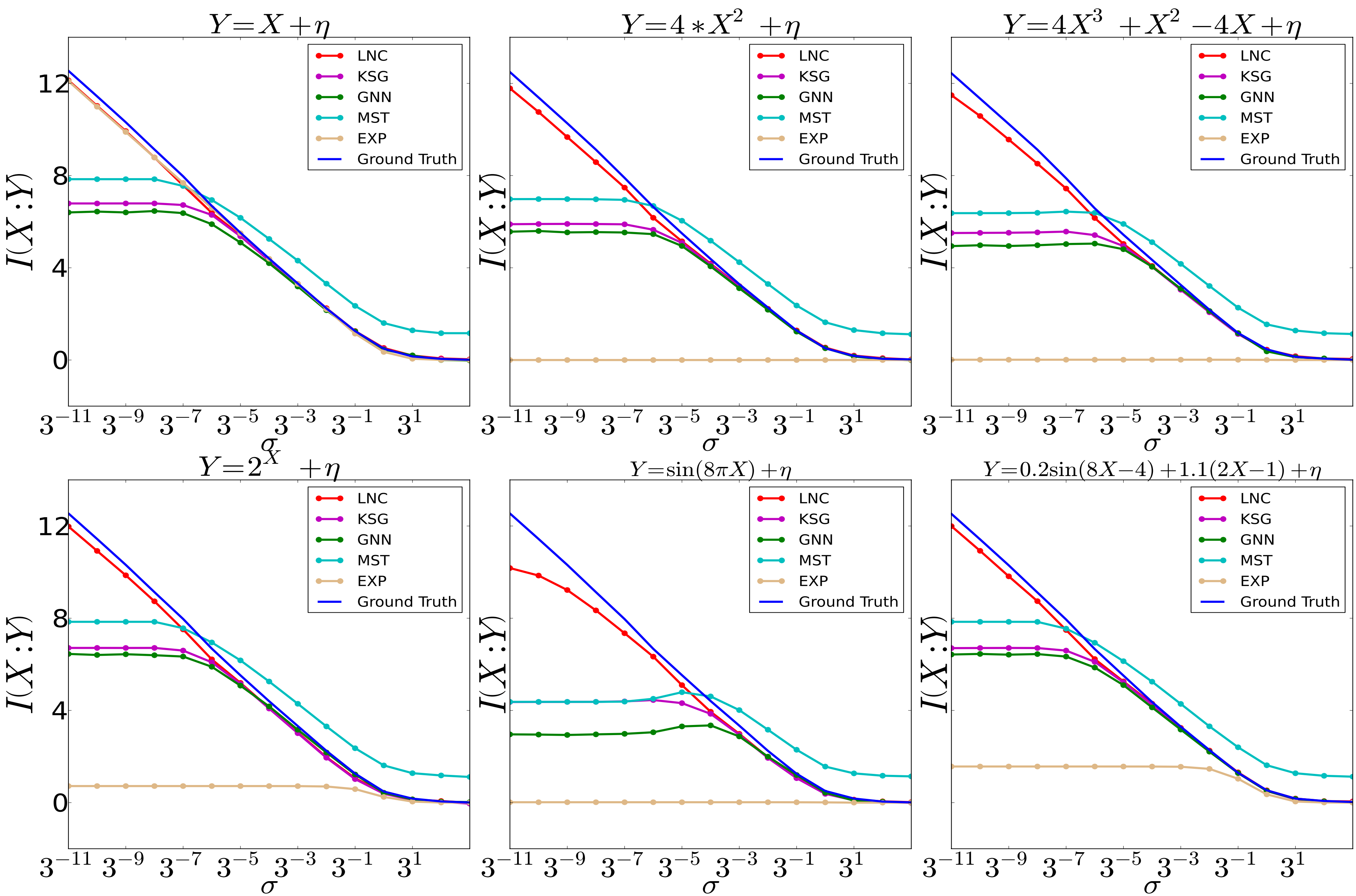
$\overline{V}(\boldsymbol{i})/V(\boldsymbol{i}) < \alpha_{k,d}$



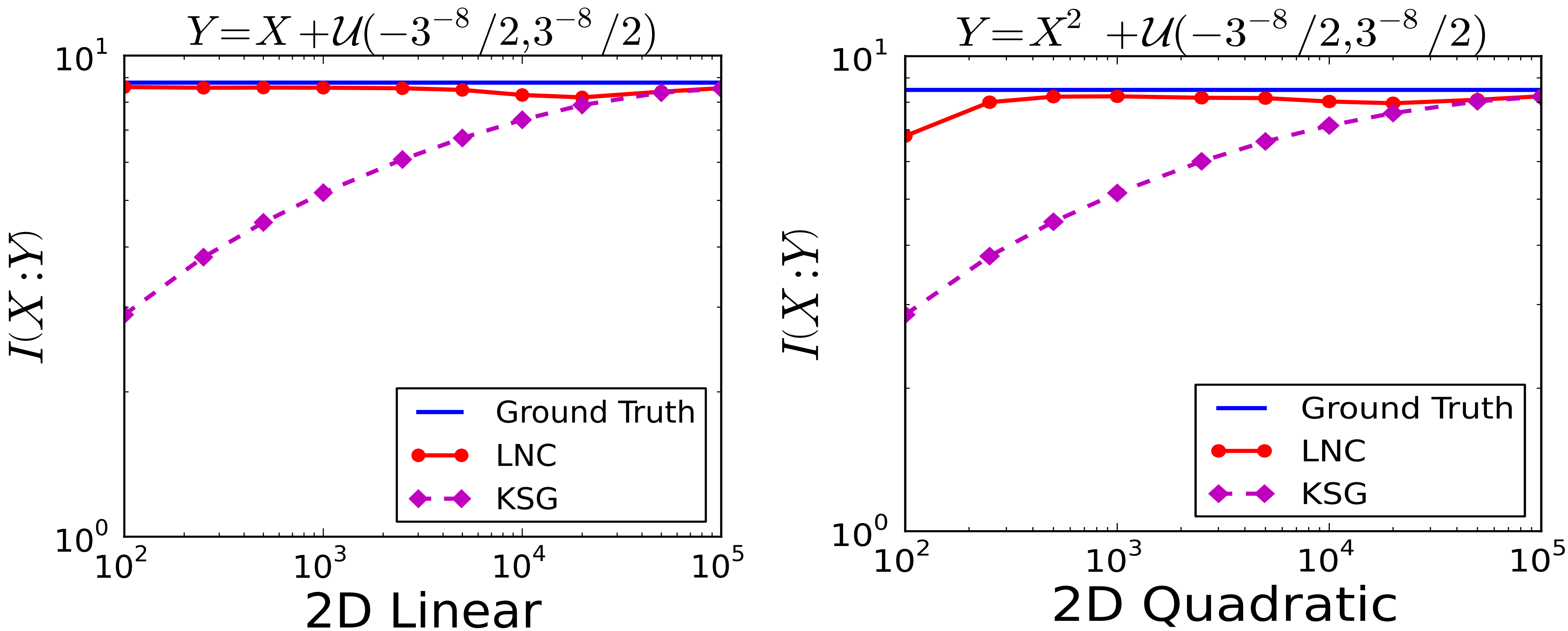
Functional Relationships



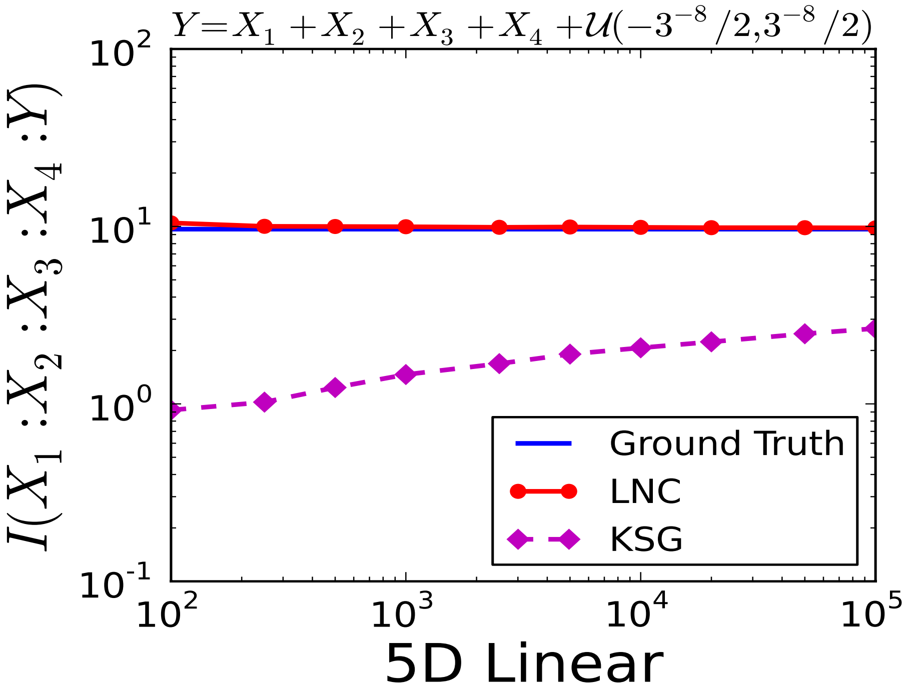
Functional Relationships

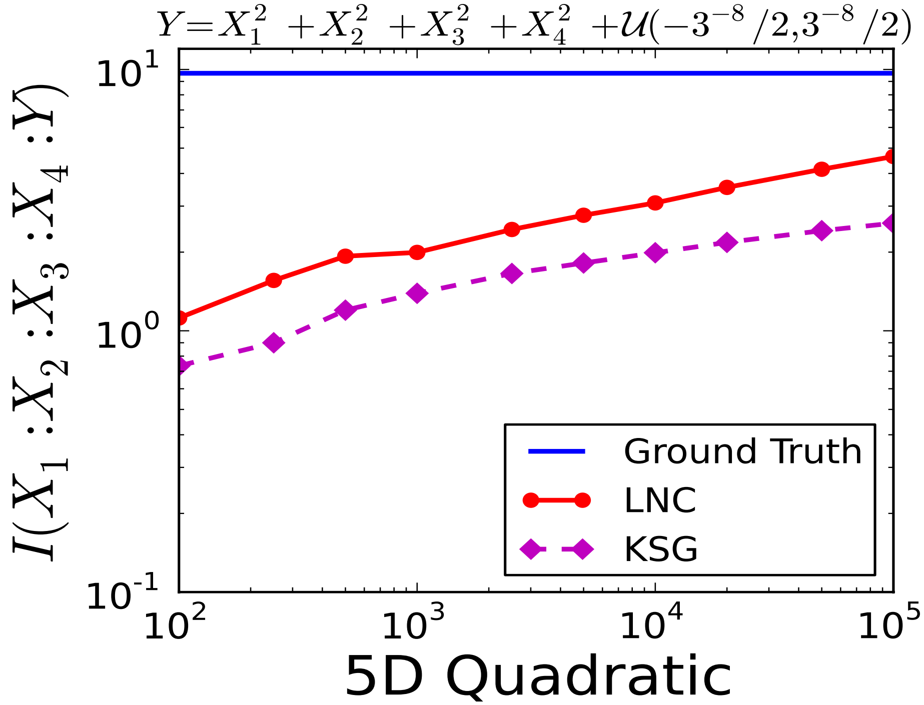


Empirical Convergence Rate

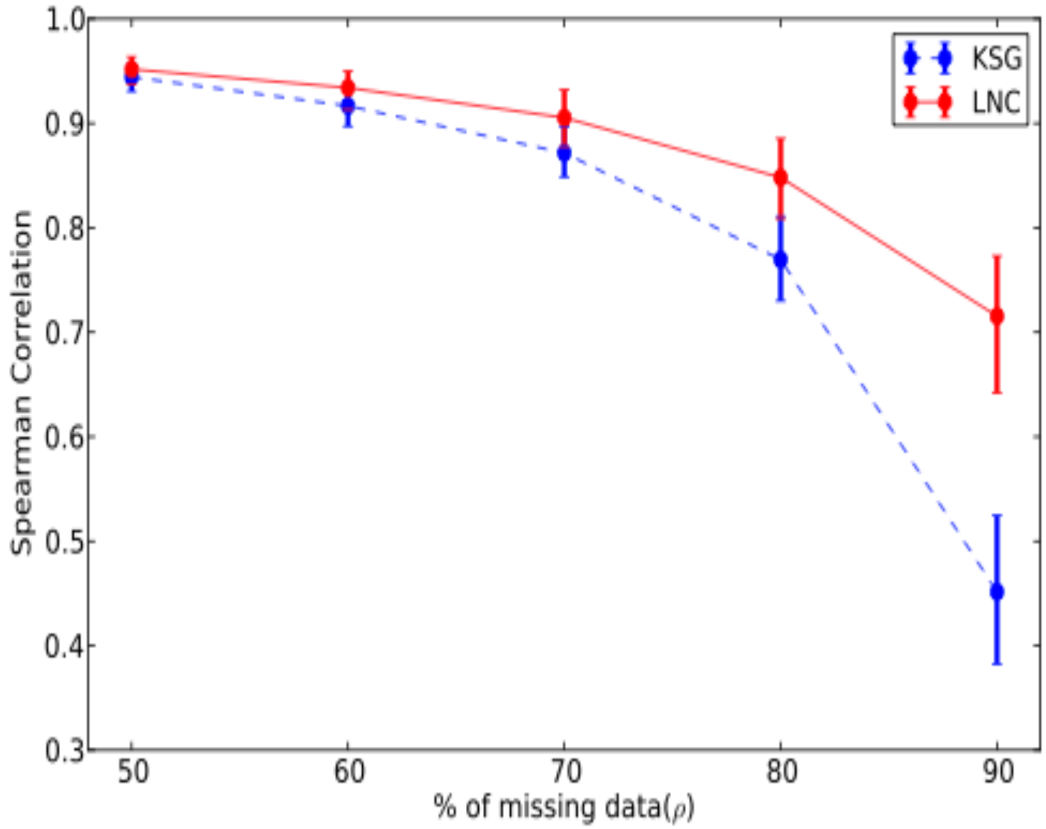


Empirical Convergence Rate





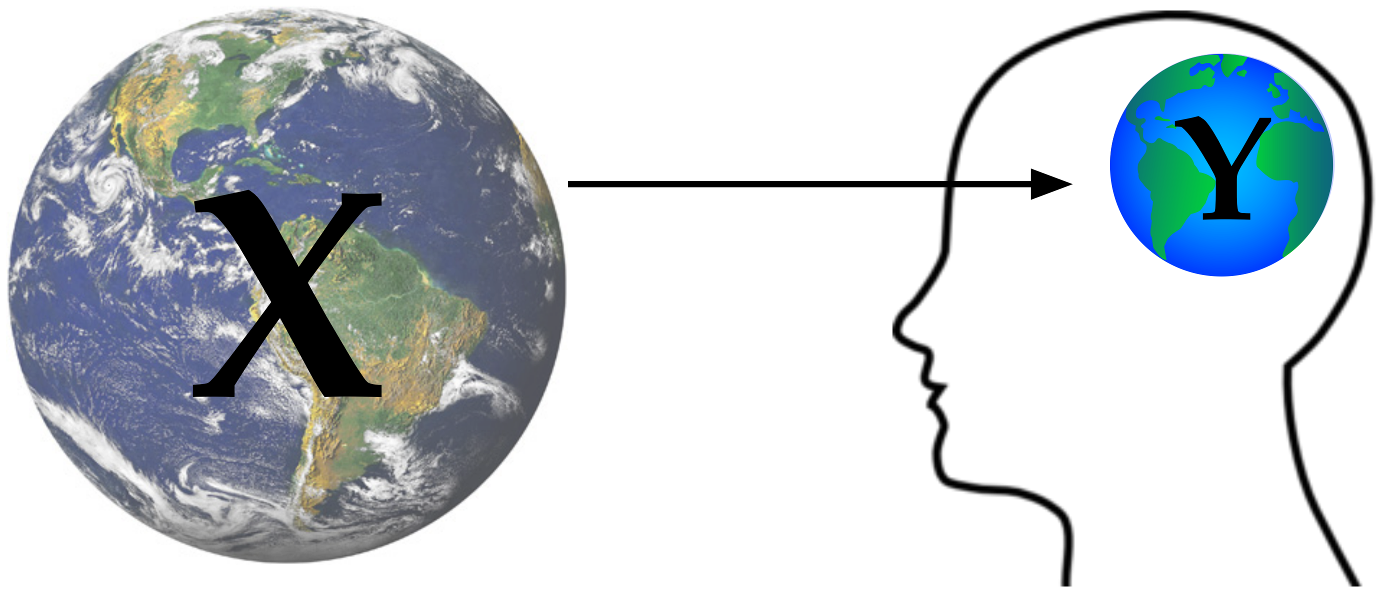
Ranking Relationship Strength



- **WHO (World Health Organization) data-set: 357 socio-economic variables**
- **We ranked the relationship strength between pairs of variables based on mutual information**
- **Tested the robustness of ranking under missing data.**

- Information Theory Basics
  - Entropy, MI, Discrete IT estimators
  - Entropy estimation demo
- Human behavior dynamics
  - Social networks
  - Stylistic coordination
- **Coffee Break (3:15-3:30)**
- Non-parametric entropy estimation
- **Very high-dimensional information**
  - How to handle it?
  - Applications: language, personality, behavior

Representing high-dimensional information



Problem

Information is a functional of p(x)

If x is "medium dimensional" then
we can use our estimation tricks.

But what if x is truly high dimensional?

Approaches

- Don’t even try (i.e., pick a low-d problem)
  - Dimensionality reduction

- **Compression**

- Information decomposition

Compression: InfoMax

$$\max_{p(y|x)} I(X;Y)$$

Mutual information is maximized if we copy the information.

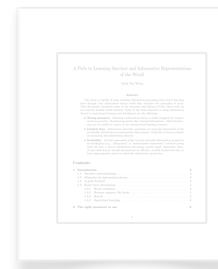
1 bit of noise = 1 bit of signal!

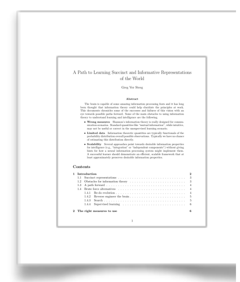
Infomax representations produce a copy of a copy of a copy…

This is really an alternate statement of the Data Processing (in)equality

$$I(X;Y^1,\ldots,Y^k) = I(X;Y^1)$$

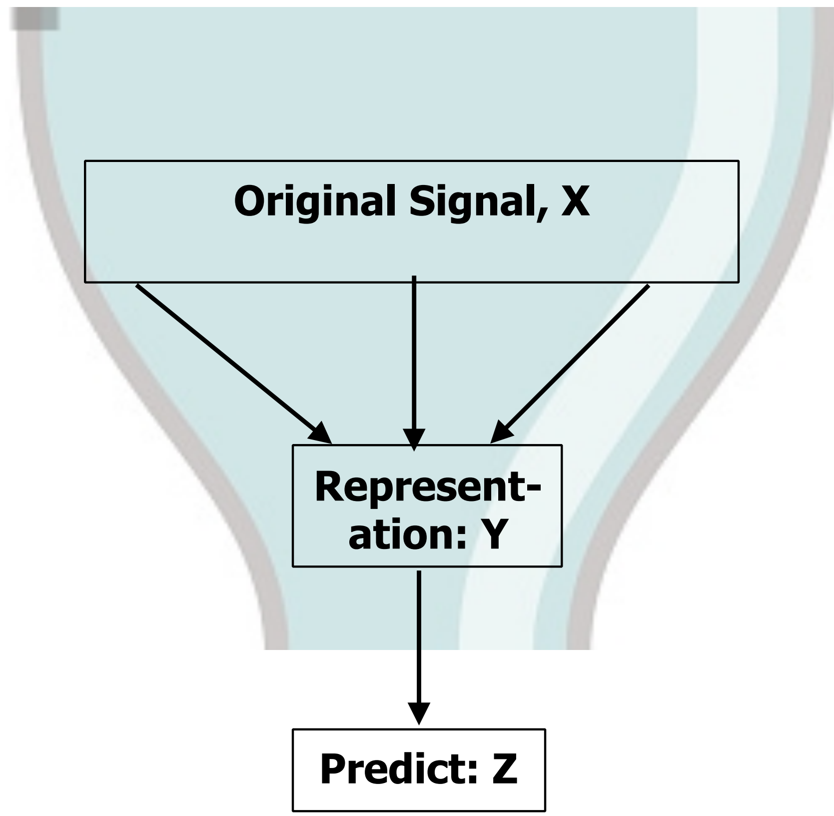
$$\begin{array}{c} Y^3 \\ \uparrow \\ Y^2 \\ \uparrow \\ Y^1 \\ \uparrow \\ X \end{array}$$





A deep representation, each symbol is a layer

Compression: the information bottleneck



Tishby, Slonim, et al.
(Rate-distortion)

$$\min_{p(y|x)} I(X;Y) - \gamma I(Y;Z)$$

Approaches

- Don’t even try (i.e., pick a low-d problem)
    - Dimensionality reduction
- Compression
- **Information decomposition**

Extending mutual information

- **Entropy** the average number of bits required to store X

$$H(X) = -\sum_x p(x) \log p(x)$$

- What if we want to store two variables?

Naive $$\#\ \text{bits} = H(X_1) + H(X_2)?$$

Holistic $$\#\ \text{bits} = H(X_1, X_2)$$

- The difference between the naive strategy and the holistic one has a special name

$$H(X_1) + H(X_2) - H(X_1, X_2)$$

$$= \underbrace{I(X_1; X_2)}_{\textbf{Mutual information}} = TC(X_1, X_2)$$

X_1 X_2

Mutual information

"Total correlation" (Watanabe, 1967) or multivariate mutual information

$$TC(X_1, \ldots, X_n) = \sum_i H(X_i) - H(X)$$

$$= D_{KL}(p(x) || \prod_i p(x_i))$$

Holistic Naive

- Useless because we don't know p(x)

Example of decomposing the dependence

$$TC(X_1, X_2, X_3, X_4)$$

$$= \mathbb{E} \log \frac{p(x_1, x_2, x_3, x_4)}{p(x_1)p(x_2)p(x_3)p(x_4)}$$

- Let’s show this graphically before looking at the problems…

A hint to get something like "hierarchical coarse-graining"

$$
\begin{tikzcd}[column sep=small]
 & & & ((X_1, X_2), (X_3, X_4)) & & & \\
 & (X_1, X_2) \arrow[urr] & & & & (X_3, X_4) \arrow[ull] & \\
X_1 \arrow[ur] & & X_2 \arrow[ul] & & X_3 \arrow[ur] & & X_4 \arrow[ul]
\end{tikzcd}
$$

$$TC((X_1, X_2), (X_3, X_4)$$

-

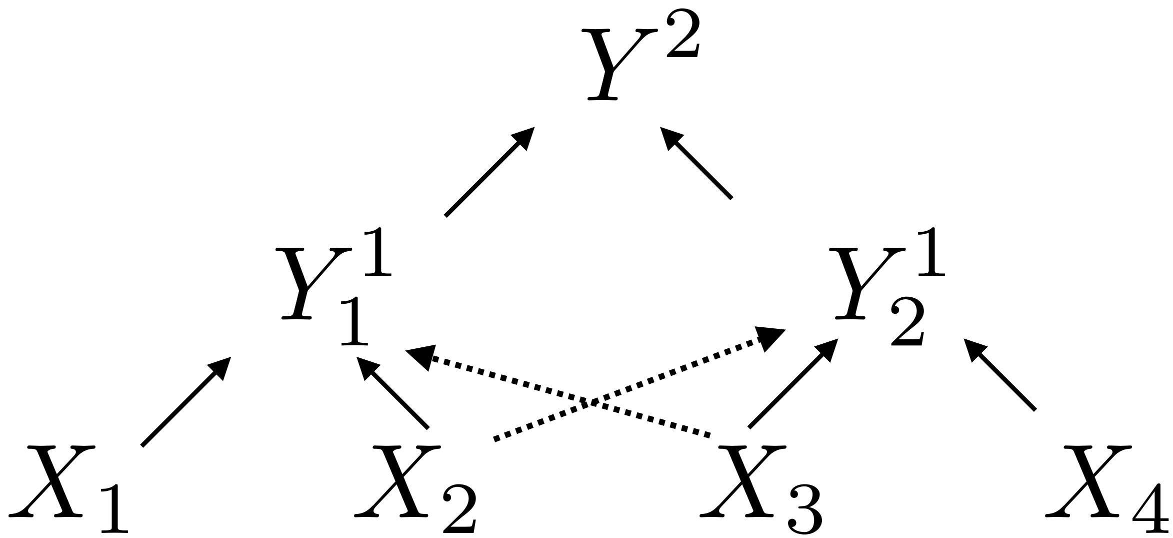
$$TC(X_1, X_2) + TC(X_3, X_4$$

=

$$TC(X_1, X_2, X_3, X_4$$

- From Watanabe's original TC paper: multivariate information can be hierarchically decomposed.
- BUT, this is only formal: it doesn't tell us the best way to decompose it, and we still get the curse of dimensionality.

A hint to get something like “hierarchical coarse-graining”



$$C(Y^1; Y^2$$

$$C(X; Y_1^1) + C(X; Y_2^1$$

$$\leq$$

$$TC(X_1, X_2, X_3, X_4$$

- Let Y’s be some arbitrary function of inputs, now we can get a lower bound
- Now optimize lower bound over functions and structure
- (An aside: Y’s at each level are more independent)

Total Correlation Explanation (CorEx)

- Total correlation or multivariate information in X

$$TC(X) \equiv D_{KL}\left(p(x)|| \prod_{i=1}^{n} p(x_i)\right)$$

- If Y were the common cause of dependence in all X_i, TC(X|Y)=0

$$TC(X|Y) \equiv D_{KL}\left(p(x|y)|| \prod_{i=1}^{n} p(x_i|y)\right)$$

- The reduction in dependence, or the "correlation explained by Y"

$$TC(X;Y) \equiv TC(X) - TC(X|Y)$$

More detail on the decomposition

$$TC(X) \geq TC(X;Y^1) = TC_L(X;Y^1) + TC(Y^1)$$

Optimize this

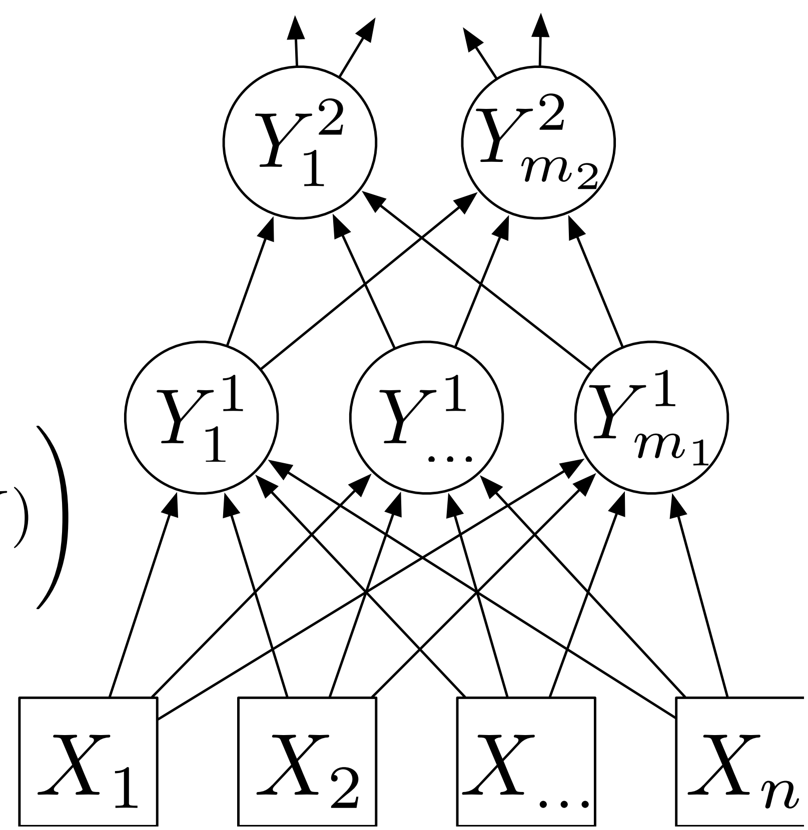
How do we get this?

$$TC(Y^1) \geq TC_L(Y^1;Y^2) + TC(Y^2)$$

...

$$TC_L(X;Y) = \sum_j \left(\sum_i \alpha_{i,j} I(X_i;Y_j) - I(Y_j : X) \right)$$

$$\alpha_{i,j} = \frac{I(X_i;Y_j|Y_{1:j-1})}{I(X_i;Y_j)}$$



Form of Solution for One Layer

$$\max_{p(y_j|x)} TC_L(X;Y)$$

Optimize over all probabilistic functions!

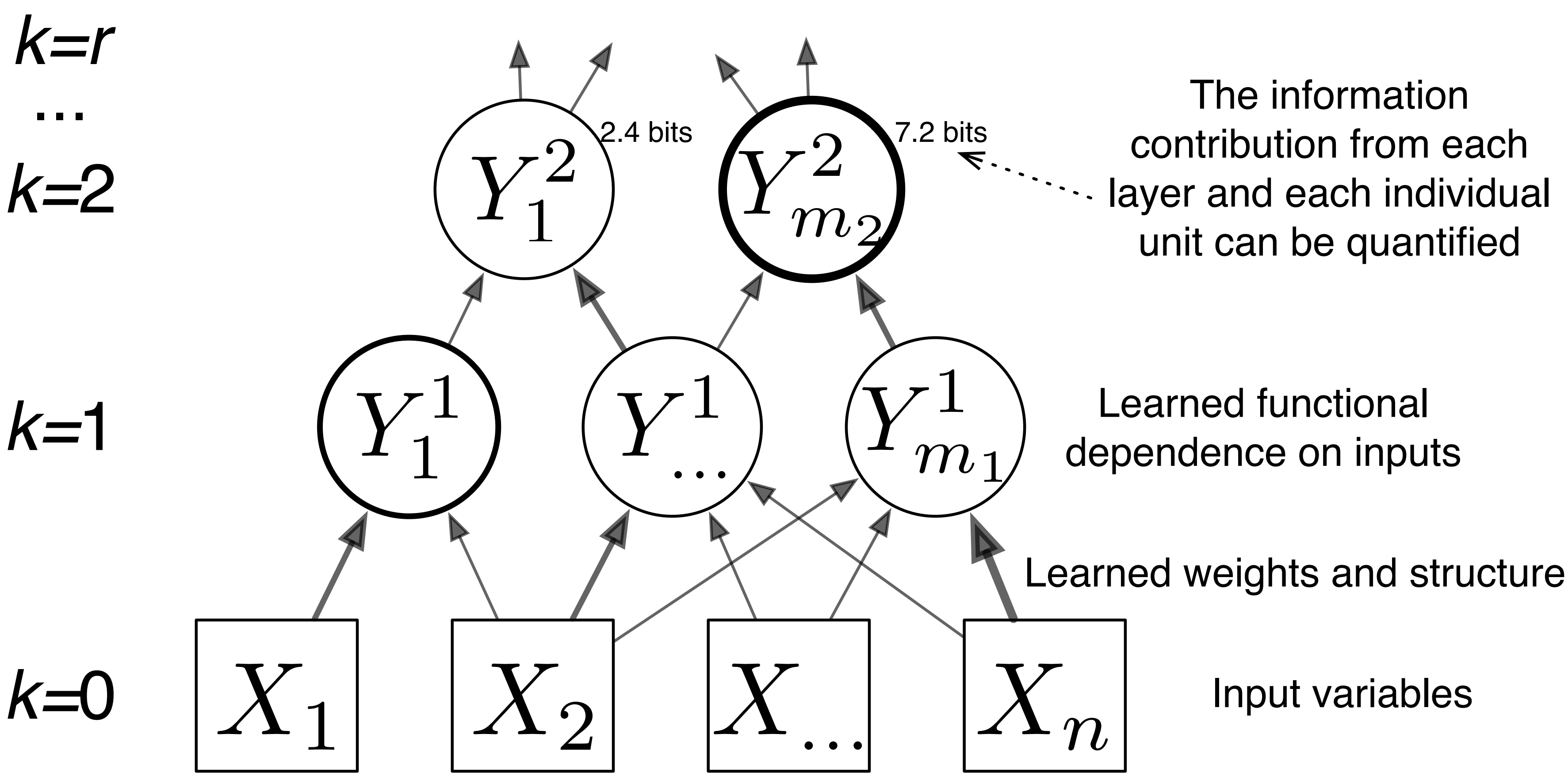
$$p(y_j|x) = \frac{p(y_j)}{Z_j(x)} \prod_{i=1}^{n} \left(\frac{p(y_j|x_i)}{p(y_j)} \right)^{\alpha_{i,j}}$$

Structure (a principled criteria naturally arises: links for "unique" info)

Z is easy to calculate and gives an estimate of the objective for free.

Depends on marginals only

What the visualizations will summarize

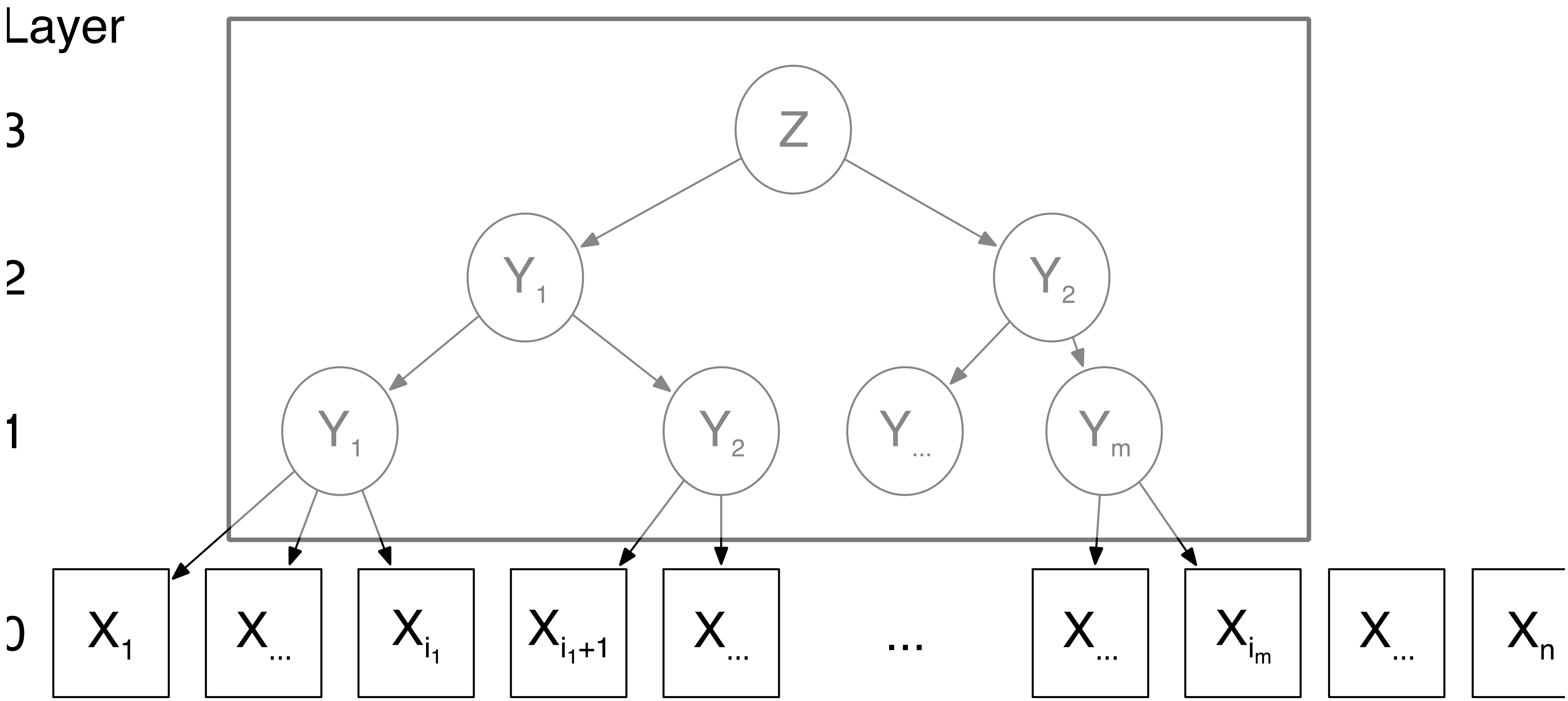


$$TC(X) \geq \sum \text{contribution from } Y_j^k$$

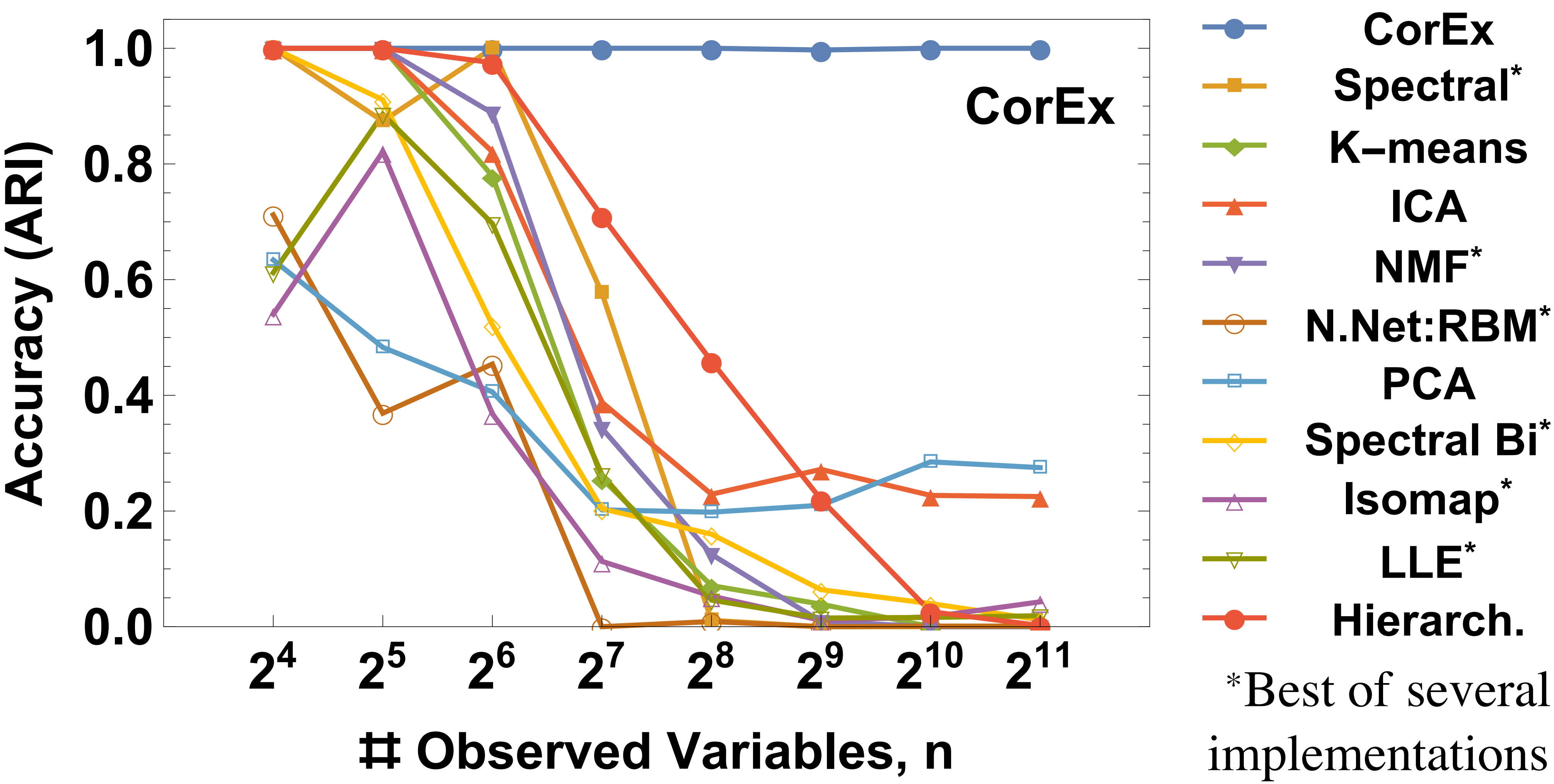
Applications

Benchmark test: Reconstruct latent tree models

Goal: recover the hidden structure generating this data



Accuracy to recover structure for high-d tree models



There are also specialized techniques dedicated to latent tree learning: the complexity of these are O(n^3) – O(n^5), none could run on these examples with thousands of variables

The Big-5 personality test

Q31: I am the life of the party

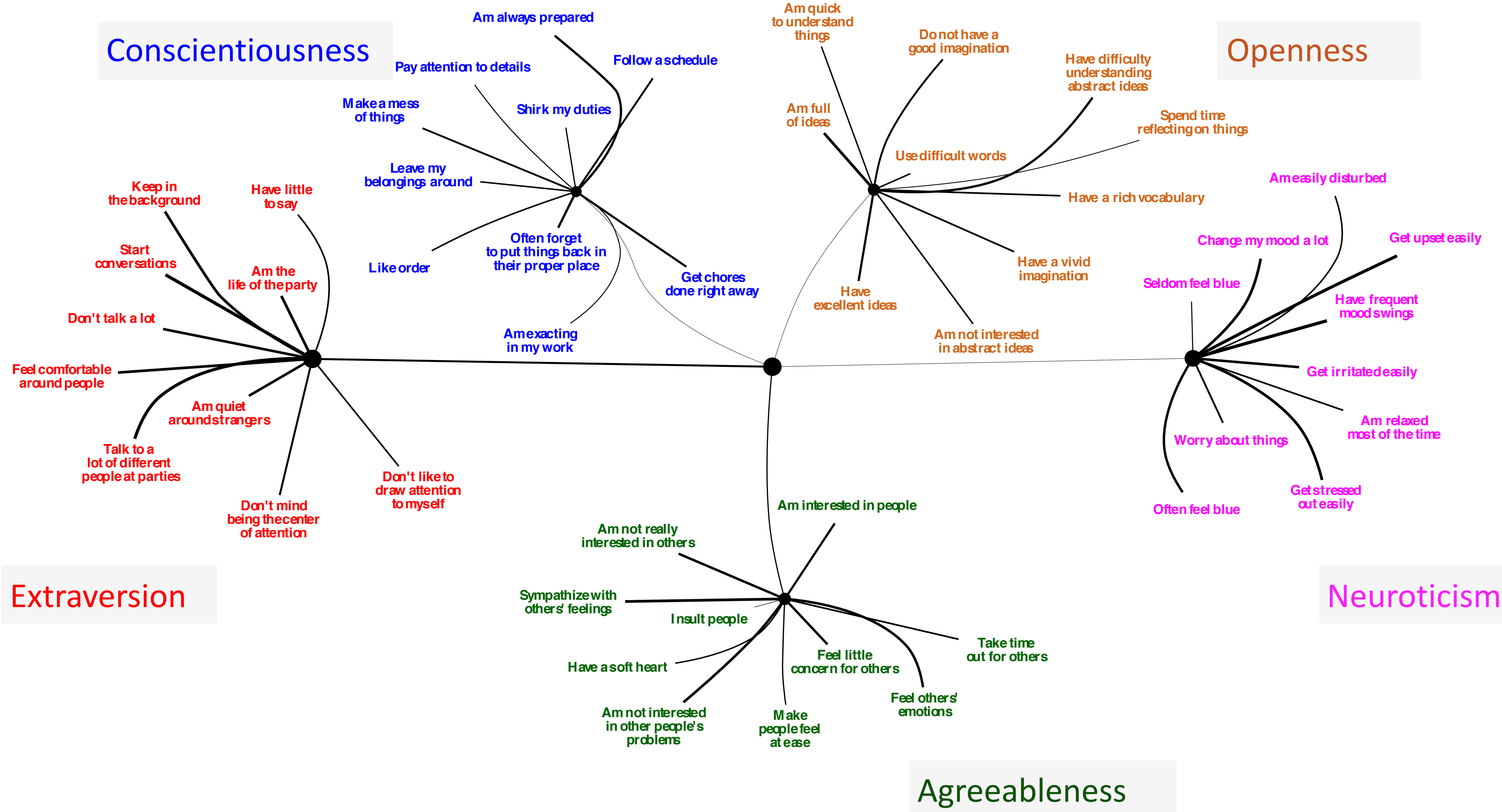
1. Strongly disagree
2. Disagree
3. Neither agree nor disagree
4. Agree
5. Strongly agree

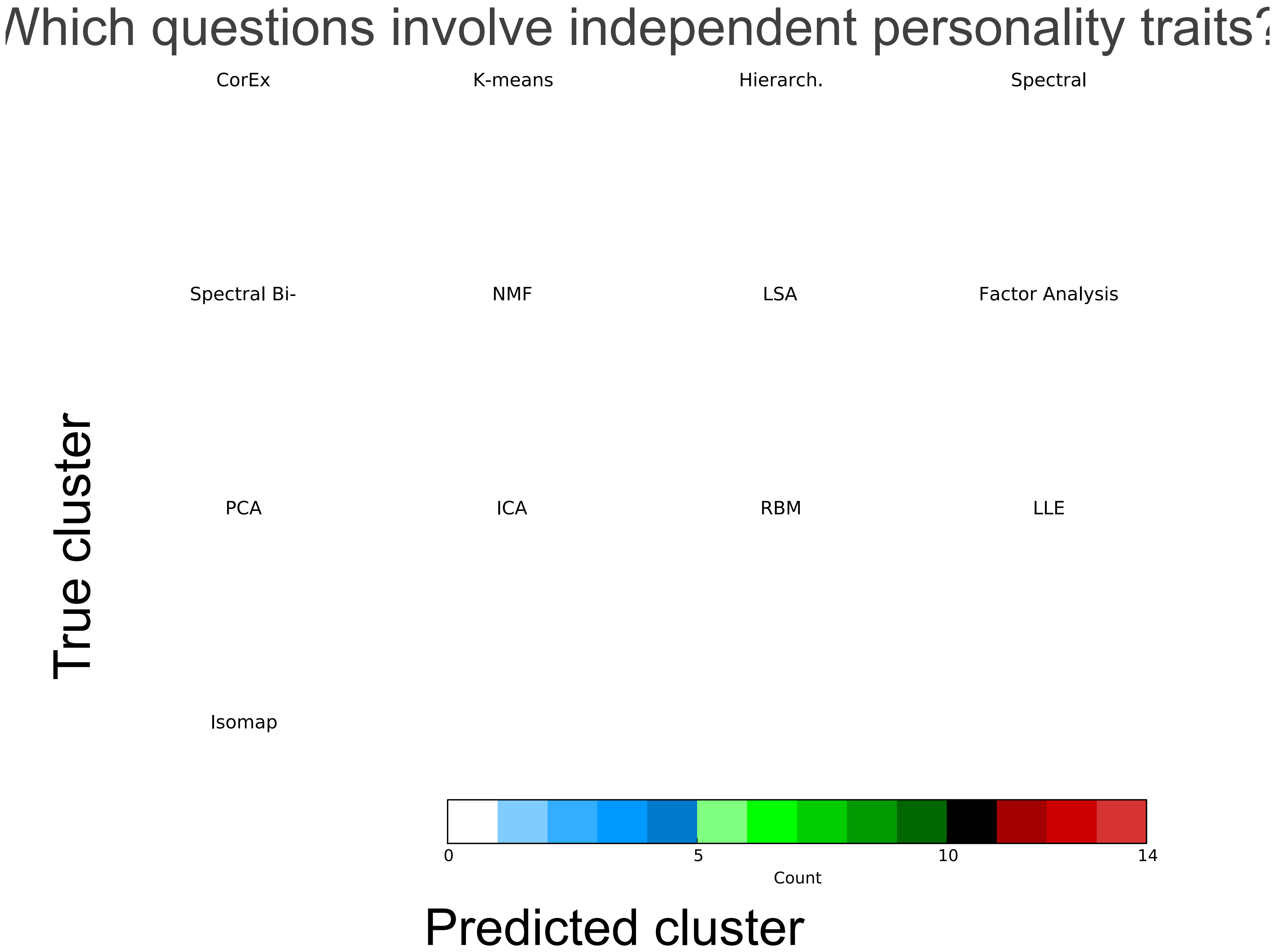
	Q1	Q2	Q3	…	Q50
Person 1	**5**	**2**	**4**		**1**
…					
Person N	**2**	**2**	**5**		**5**

According to psychologists, this question measures **Extroversion**, one of the “Big 5” personality traits.

Given answers to many questions, can we reverse engineer personality types?

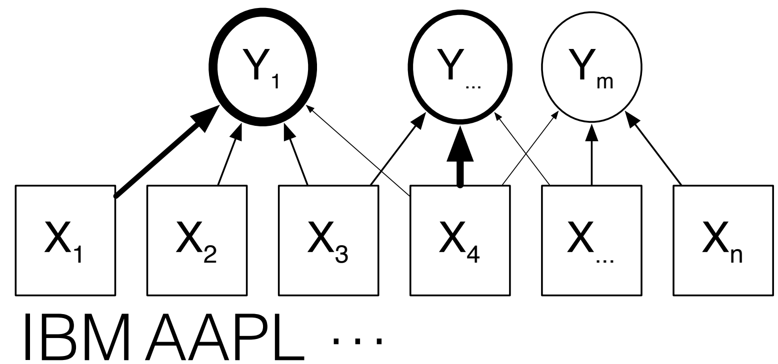
Perfect Recovery of "Big 5" Personality Traits from Survey Data




Which questions involve independent personality traits?
CorEx
K-means
Hierarch.
Spectral
Spectral Bi-
NMF
LSA
Factor Analysis
PCA
ICA
RBM
LLE
Isomap
True cluster
0
5
10
14
Count
Predicted cluster

Individual trading behavior

- Each variable represents whether an individual trades on a certain company (in a 6 month time-frame)
- Each account's activity is a sample



Grain of salt: Experiment restricted to frequent traders and frequently traded stocks

X = (1, 2, 0, 0, …)
I bought IBM, sold AAPL in this time period

[Some slides removed]

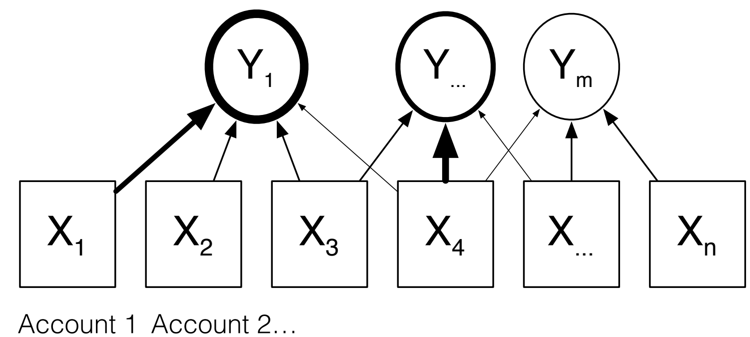
Dynamics

- Considered just one stock: **AAPL**

- 110 trading days from:
  Jan.2 2014 - Jun. 10 2014

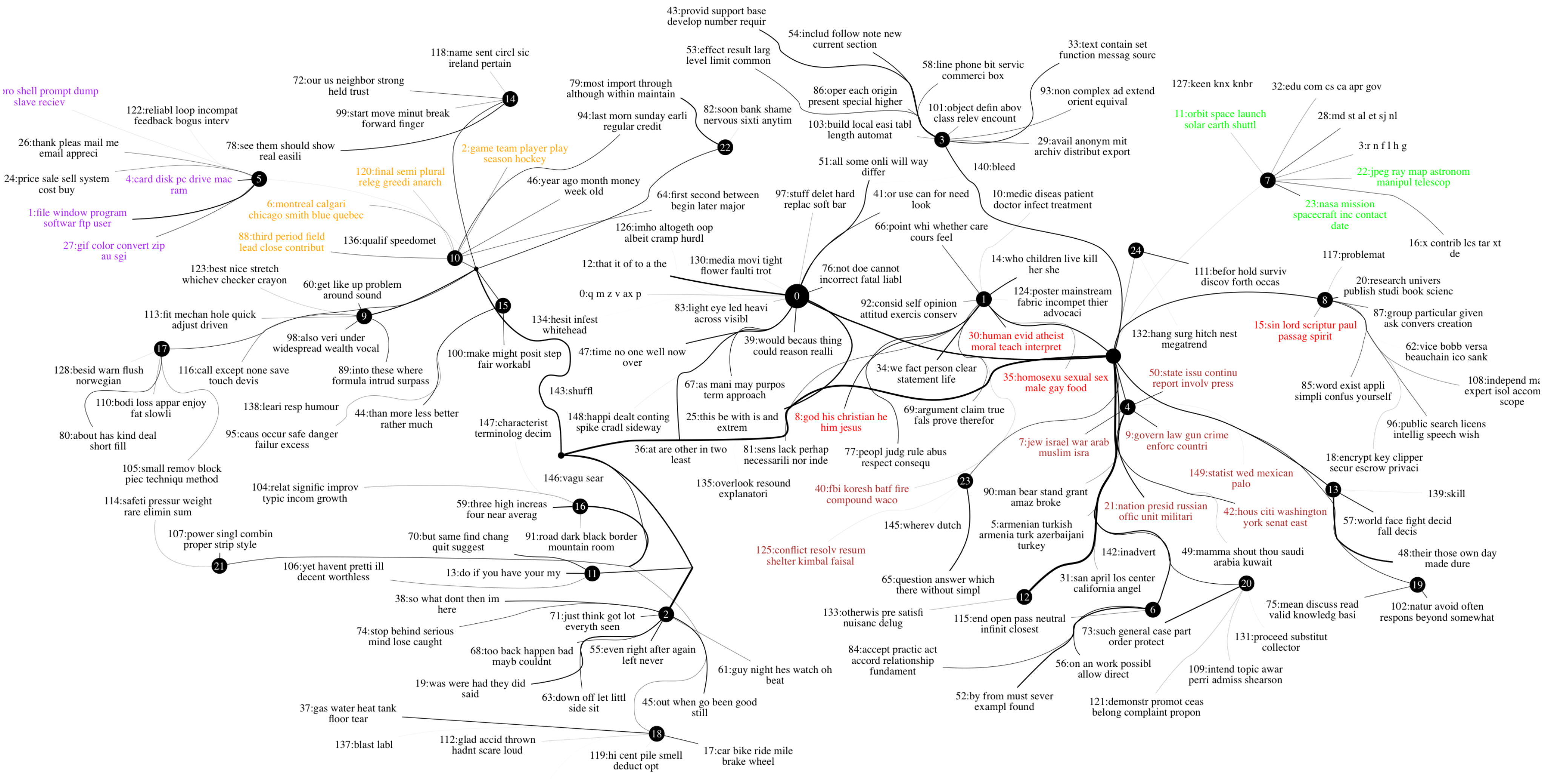
- Each day represents a sample of activity

- Variables are accounts, indicate buy/sell/both/neither for that day



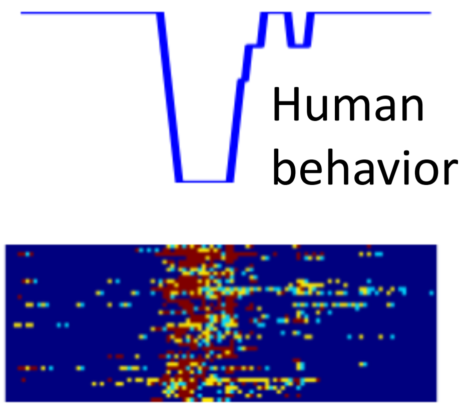
Application to hierarchical topic modeling

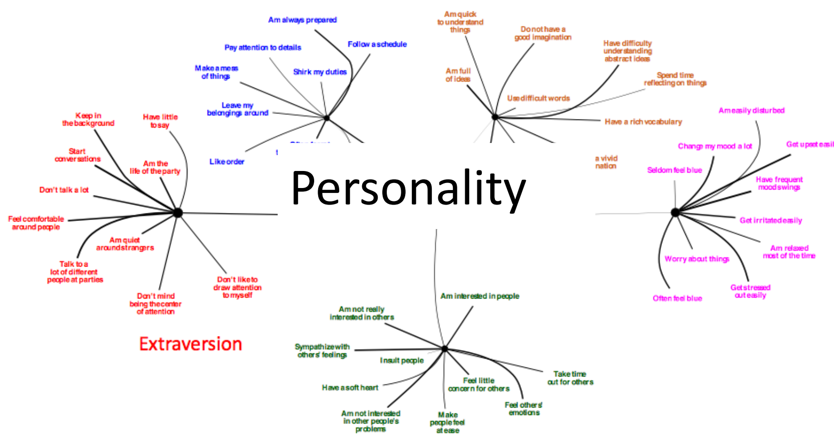
- Data from 20 newsgroups
- Each document is a sample, each word is a variable
- Hierarchical decomposition:

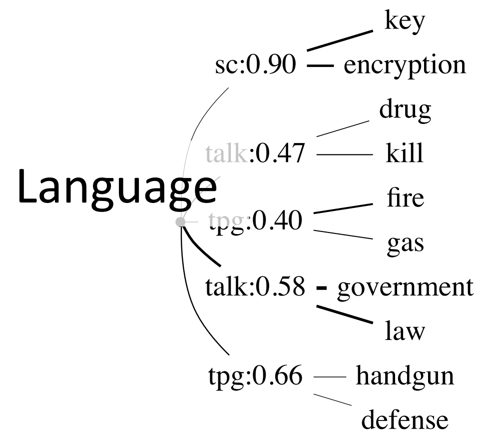


Zooming in on some example results

CorEx wrap-up







- Promising: an information-theoretic path to create succinct representations of complex data in an unsupervised way
- Practical: works on *high-d data* with *few samples* and *no assumptions* about data-generating process

Contact: gregv@isi.edu, galstyan@isi.edu

Papers, open source code, interactive visualizations: http://bit.ly/corex_info

Overall wrap-up

- Information theory is a general but challenging way to measure the strength of relationships
- We use this in hard to model domains, like social network dynamics
- For medium or low-dimensional problems, careful estimation solves most of our problems
- For very high-dimensional systems, we can use information decomposition (CorEx)

Contact: gregv@isi.edu, galstyan@isi.edu

ICWSM Tutorial: http://isi.edu/~galstyan/icwsm13

CorEx: http://bit.ly/corex_info

Entropy estimators: http://github.com/gregversteeg/NPEET

References to our related work

Social network dynamics

- Greg Ver Steeg and Aram Galstyan. Information transfer in social media. In Proceedings of World Wide Web Conference (WWW), 2012.
- Greg Ver Steeg and Aram Galstyan. Information-theoretic measures of influence based on content dynamics. In Proceedings of the 6th International Conference on Web Search and Data Mining (WSDM). ACM, 2013.
-

CorEx

- Greg Ver Steeg and Aram Galstyan. Discovering structure in high-dimensional data through correlation explanation. Advances in Neural Information Processing Systems (NIPS), 2014.
- Greg Ver Steeg and Aram Galstyan. Maximally informative hierarchical representations of high-dimensional data. In Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics (AISTATS), 2015.
-

Stylistic accommodation

- Shuyang Gao, Greg Ver Steeg, and Aram Galstyan. Understanding confounding effects in linguistic coordination: an information-theoretic approach. PLoS ONE, 10(6): e0130167, 2015.

- **Entropy estimation**
- Shuyang Gao, Greg Ver Steeg, and Aram Galstyan. Estimating mutual information by local gaussian approximation. In Uncertainty in Artificial Intelligence (UAI), 2015.
- Shuyang Gao, Greg Ver Steeg, and Aram Galstyan. Efficient estimation of mutual information for strongly dependent variables. In Proceedings of the Sixteenth International Conference on Artificial