Information-Theoretic Tools for Social Media

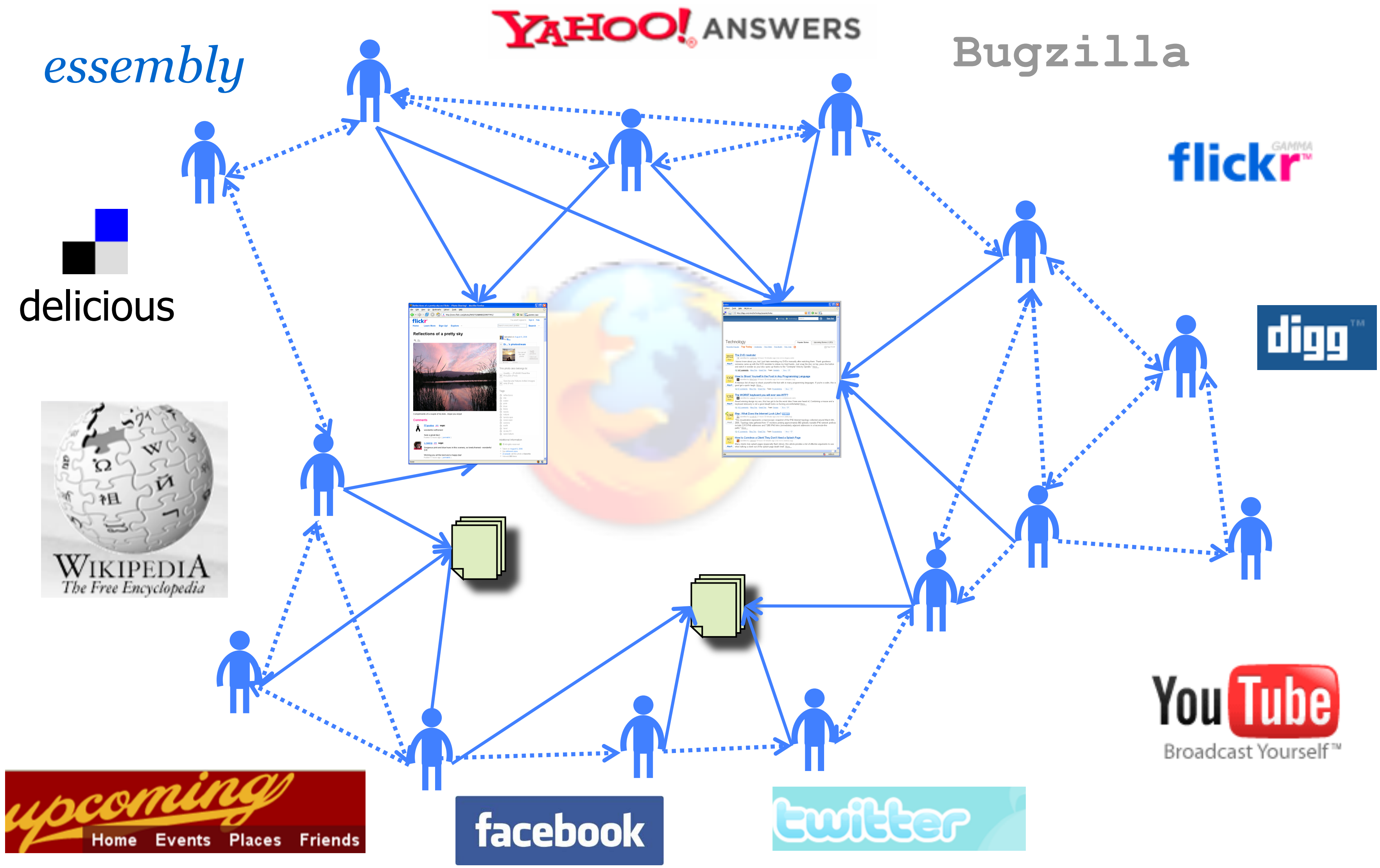
Aram Galstyan
USC Information Sciences Institute

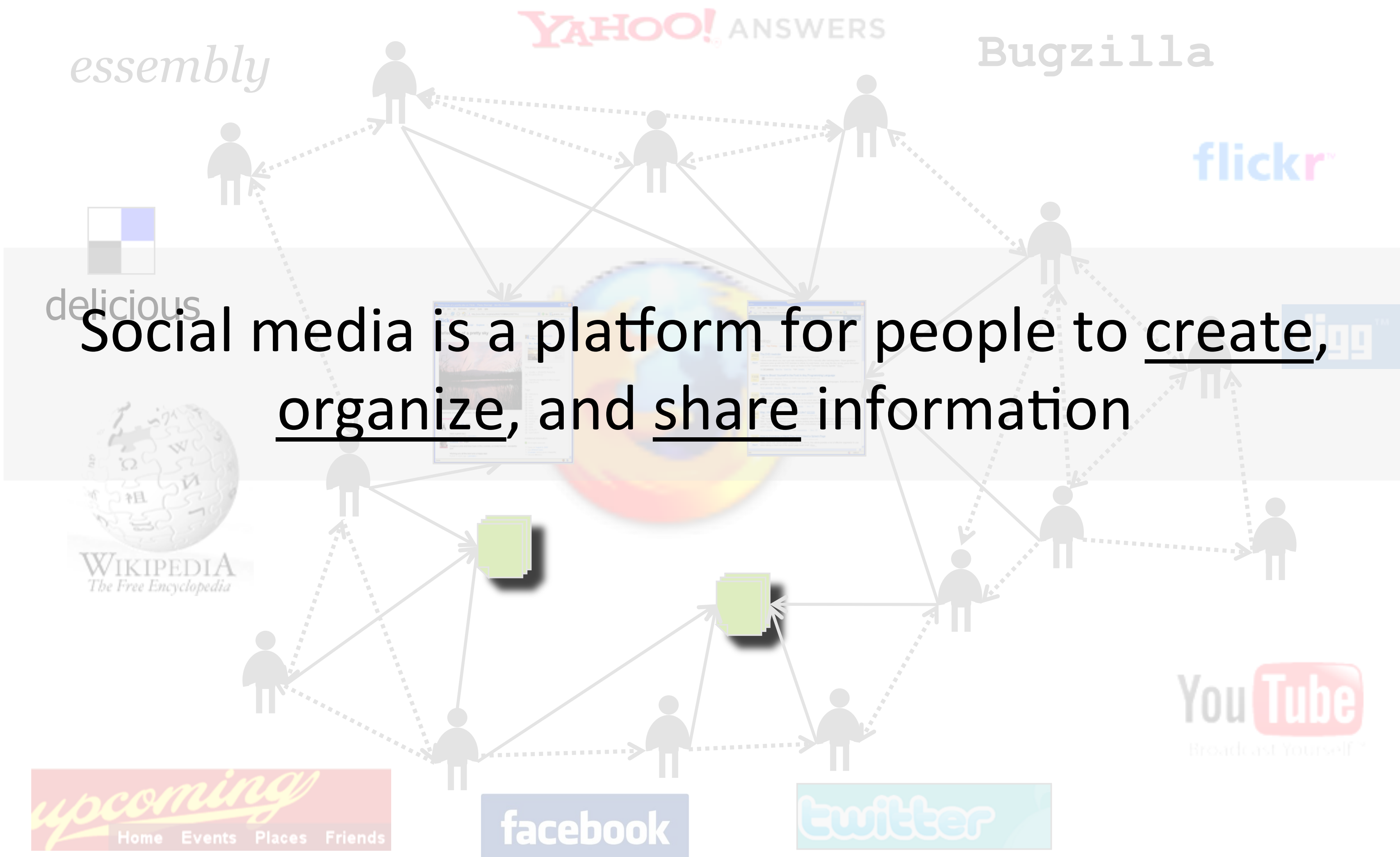
joint work with Greg Ver Steeg

USC VITERBI SCHOOL OF ENGINEERING

UCSD
March 5, 2013

ISI Information Sciences Institute


essembly
YAHOO! ANSWERS
Bugzilla
flickr
delicious
digg
WIKIPEDIA
The Free Encyclopedia
YouTube
Broadcast Yourself™
upcoming
Home Events Places Friends
facebook
twitter



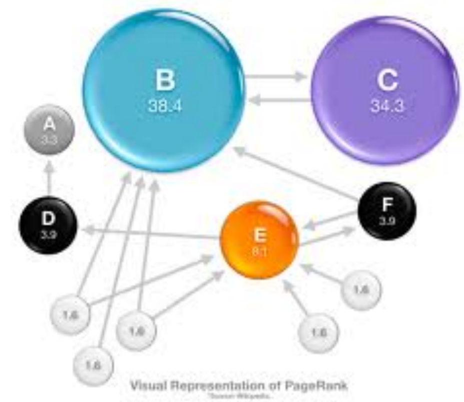
Social media is a platform for people to create, organize, and share information

Research Problems

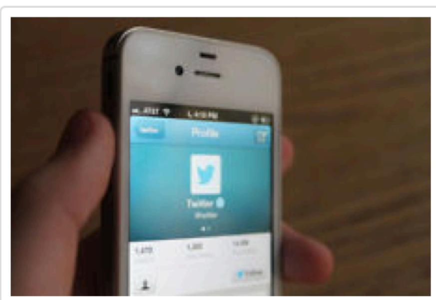
- How social networks form and change with time?
  - Network growth models
- How information flows through social networks?
  - Impact of network structure on information diffusion
- What topics are discussed and how do they evolve?
  - Detecting trending topics & real-world events
- How to find influential nodes in the network?
  - How to characterize *influence*?

Measuring influence

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



Twitter black market on ebay



22,000 Twitter Followers Under 85 Hours No Password Required Social

One-day shipping available

25d 18h left
3/30, 3PM

$13.00
Buy It Now

Free shipping

Twitter Page with 37k+ followers

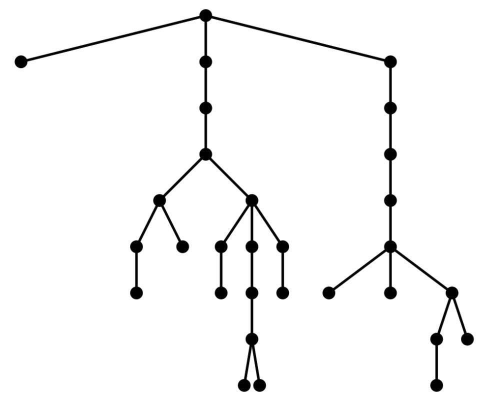
42m left
Today 8:17PM

$16.00
17 bids

Free shipping

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*

Influence via Predictability

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

$$Y \longrightarrow X$$

- Quantified using *Transfer Entropy*
  - How much our uncertainty about user X's future activity is reduced by knowing Y's past activity

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

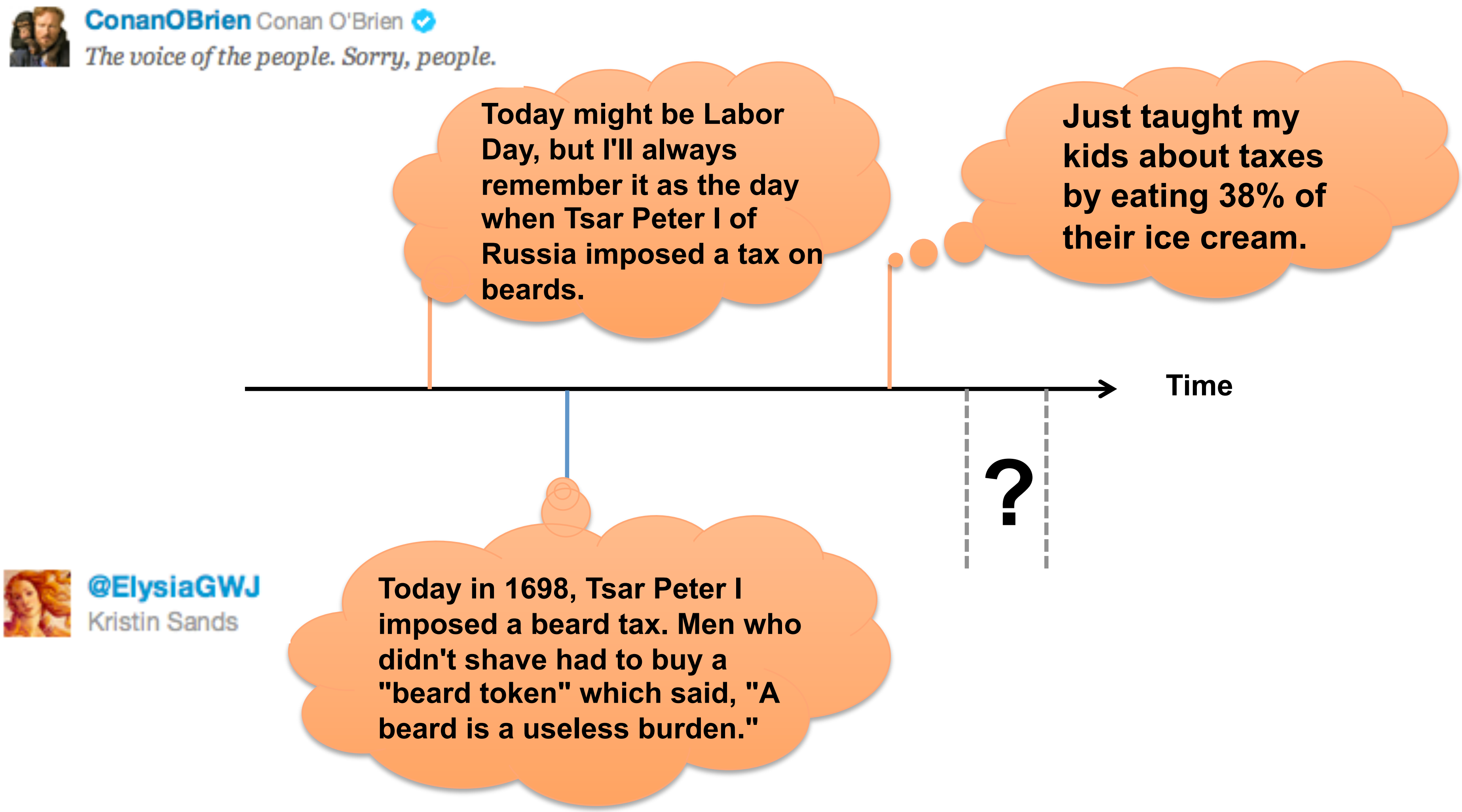
Uncertainty about X

Uncertainty about X, if you know Y's behavior

Model-free

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

Defined:
(Schrieber, PRL 85, 2000)
Related to Granger causality:
(Barnett et al, PRL 103, 2009)
Actual causality:
(Runge et al, PRL 108, 2012)


ConanOBrien Conan O'Brien
The voice of the people. Sorry, people.
Today might be Labor Day, but I'll always remember it as the day when Tsar Peter I of Russia imposed a tax on beards.
Just taught my kids about taxes by eating 38% of their ice cream.
Time
?
@ElysiaGWJ
Kristin Sands
Today in 1698, Tsar Peter I imposed a beard tax. Men who didn't shave had to buy a "beard token" which said, "A beard is a useless burden."

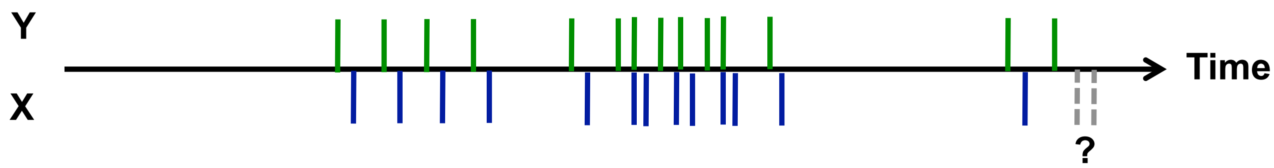
- Timing of Activity
- Content Dynamics
- Estimation of entropic measures (from limited data)

- **Timing of Activity**
- Content Dynamics
- Estimation of entropic measures (from limited data)

Transfer Entropy with Tweet Times

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

And if we add Y's history?

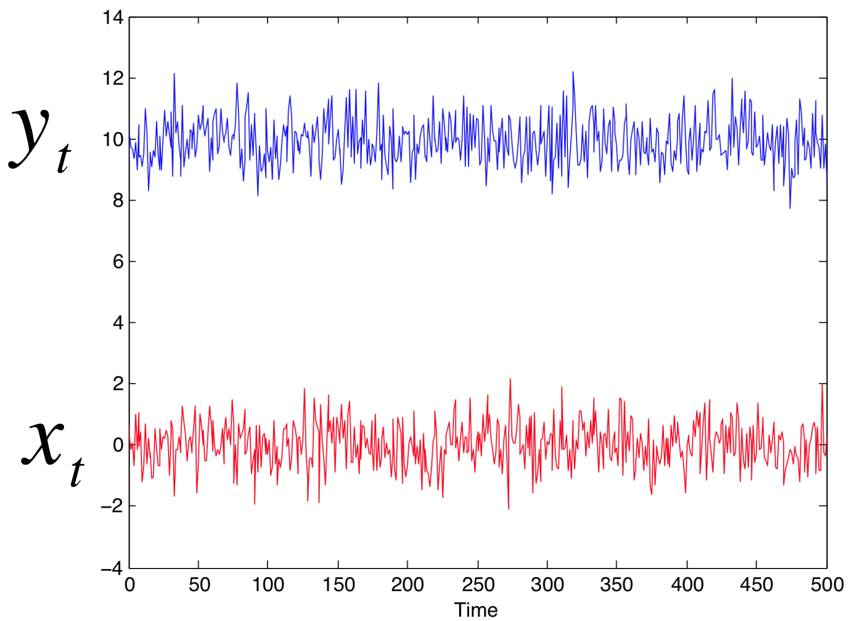


$$TE_{Y \to X} = H(X^{\mathrm{Future}} | X^{\mathrm{Past}}) - H(X^{\mathrm{Future}} | Y^{\mathrm{Past}}, X^{\mathrm{Past}})$$

Uncertainty about X

Uncertainty about X, if you know Y's behavior

Granger Causality



2003 Nobel Prize in Economics

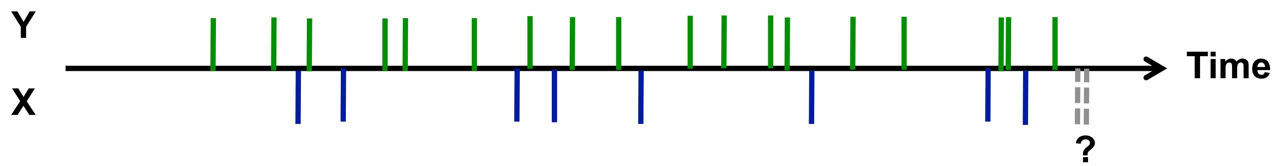
Model-1 $$x_{t+1} \approx \sum_{j=1}^{p} A_j x_{t-j}$$

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

More intuition about T.E.

Alternate possibility: low transfer entropy



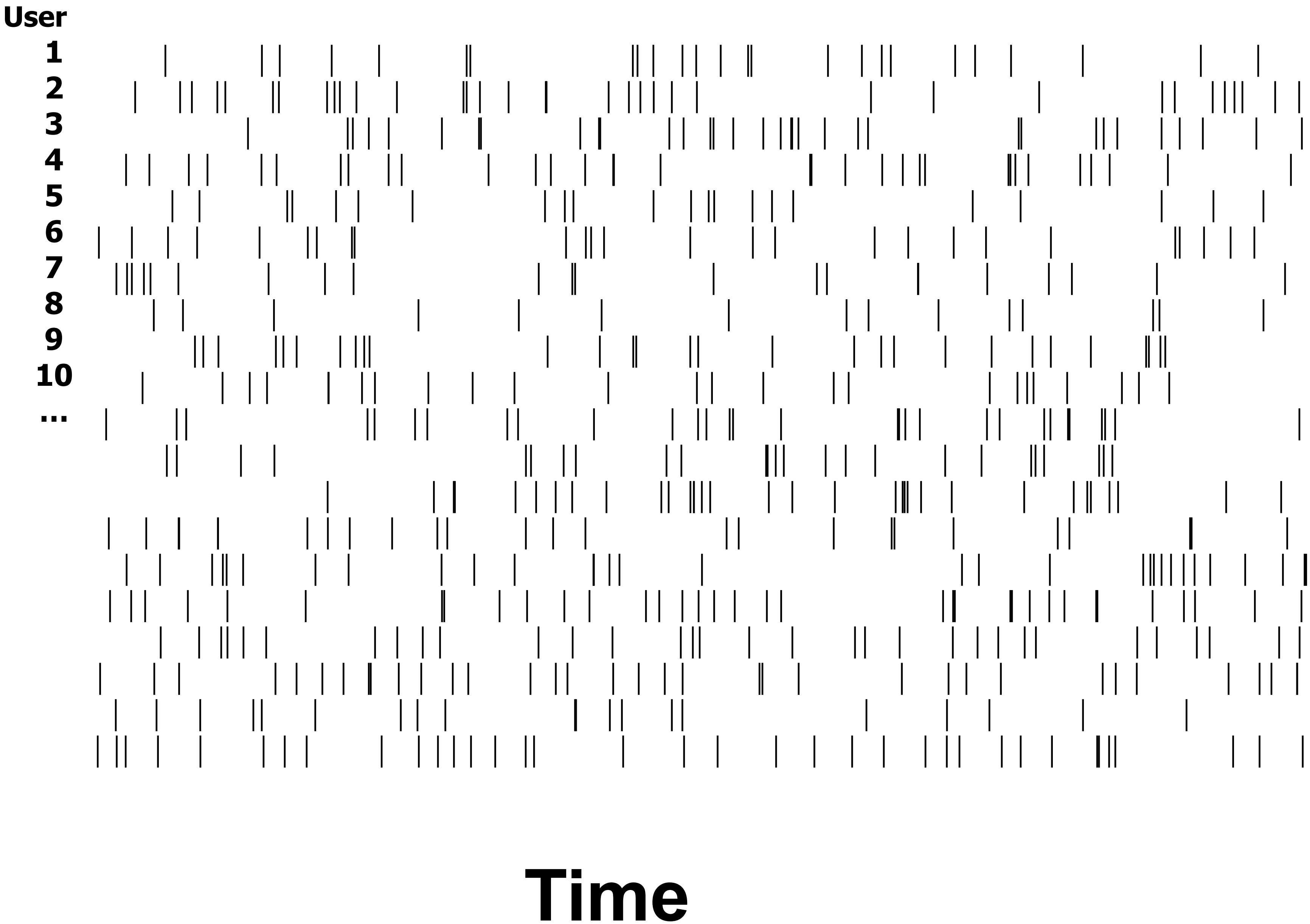
Both
Both
Random
Deterministic

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

Uncertainty about X

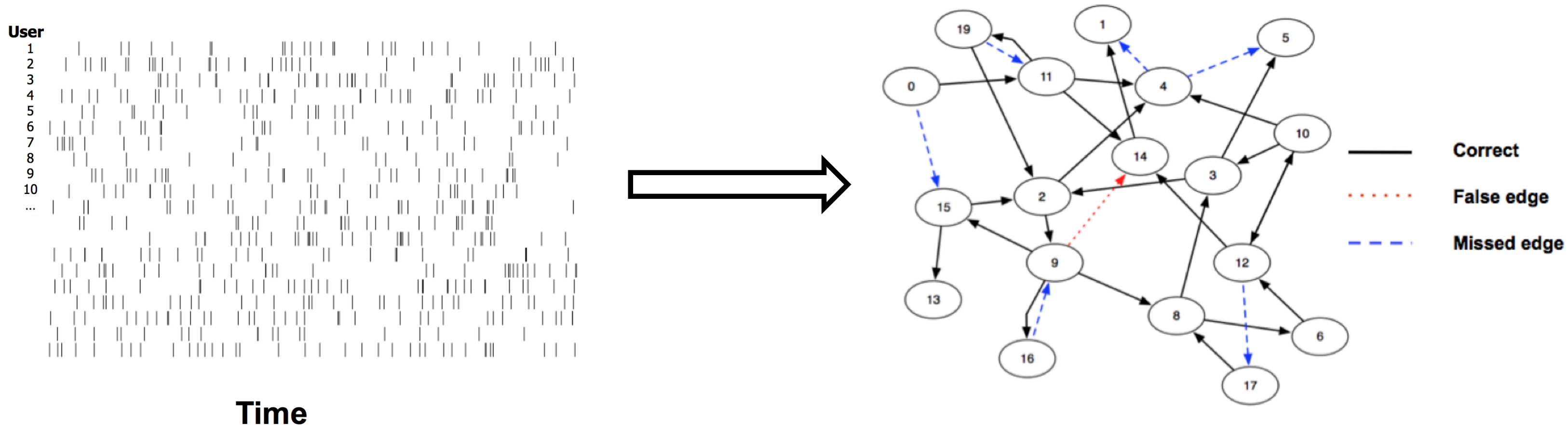
Uncertainty about X, if you know Y's behavior

Transfer entropy for tweet timing



Sample results

For synthetic model:
~ 50 posts/person for perfect reconstruction of network.



Predictable activity patterns:
- Spammers
- Political campaigns
- Fans (Bieber, etc.)
- Followback services…

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

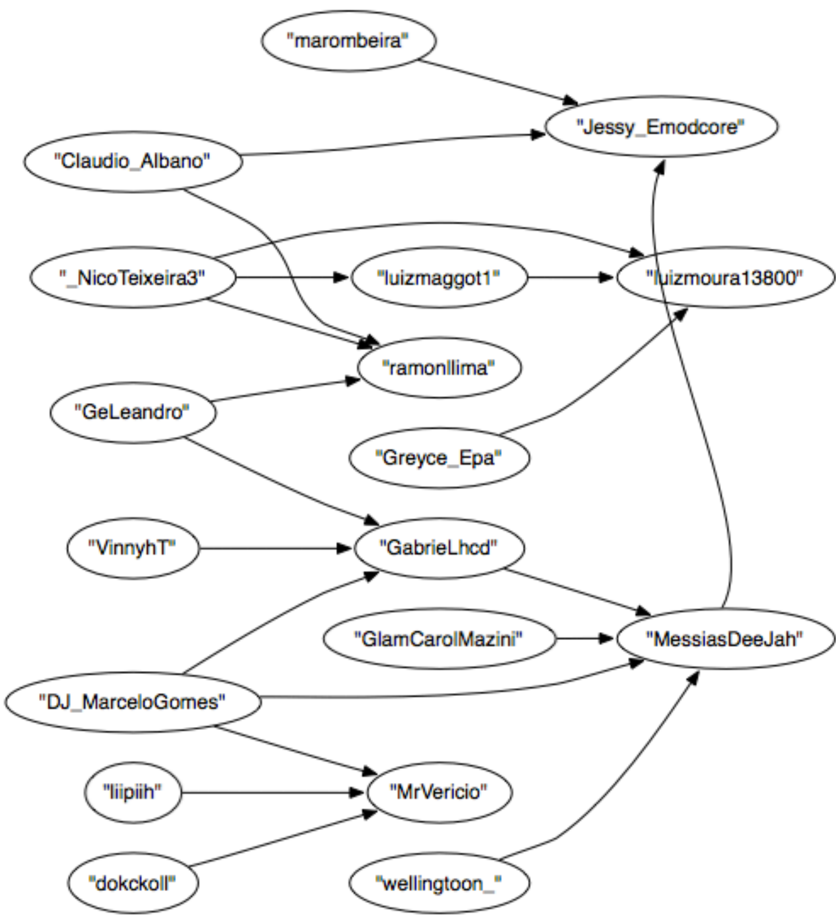
Free2BurnMusic: "#Nowplaying Janet Jackson - Hot 100 1990 http://free2burn.com/index.php #Music #IFollowBack #Music"

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

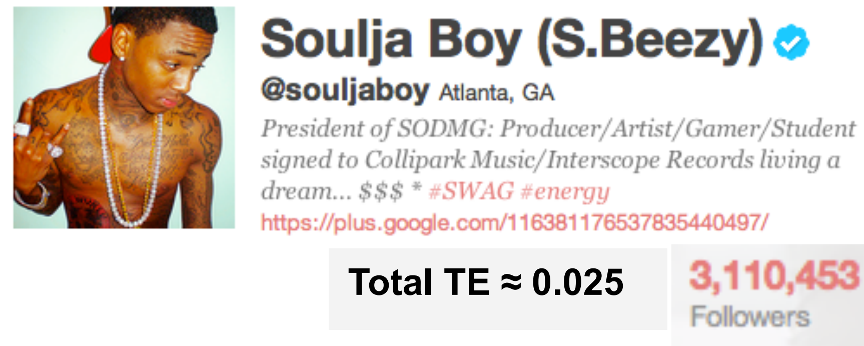
Google Translate:
Pump up your Twitter, get thousands of new followers, link to this site: http://?????? #IWantFollowers NNN

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

Tweeted over 50,000 times.

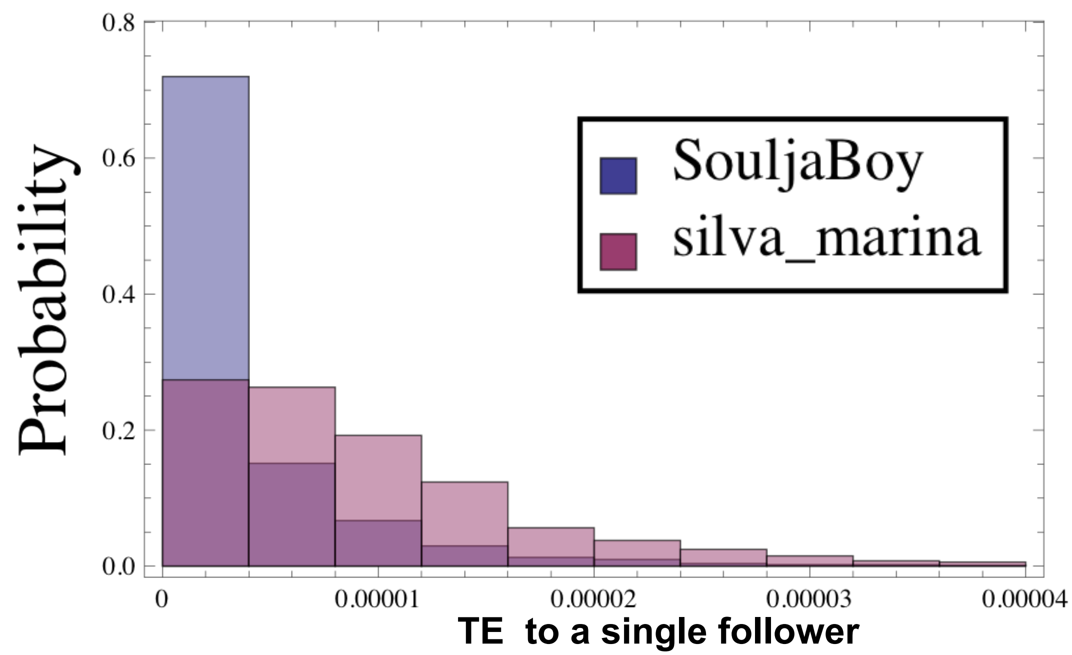
Two users with same TE

Total TE ≈ 0.025



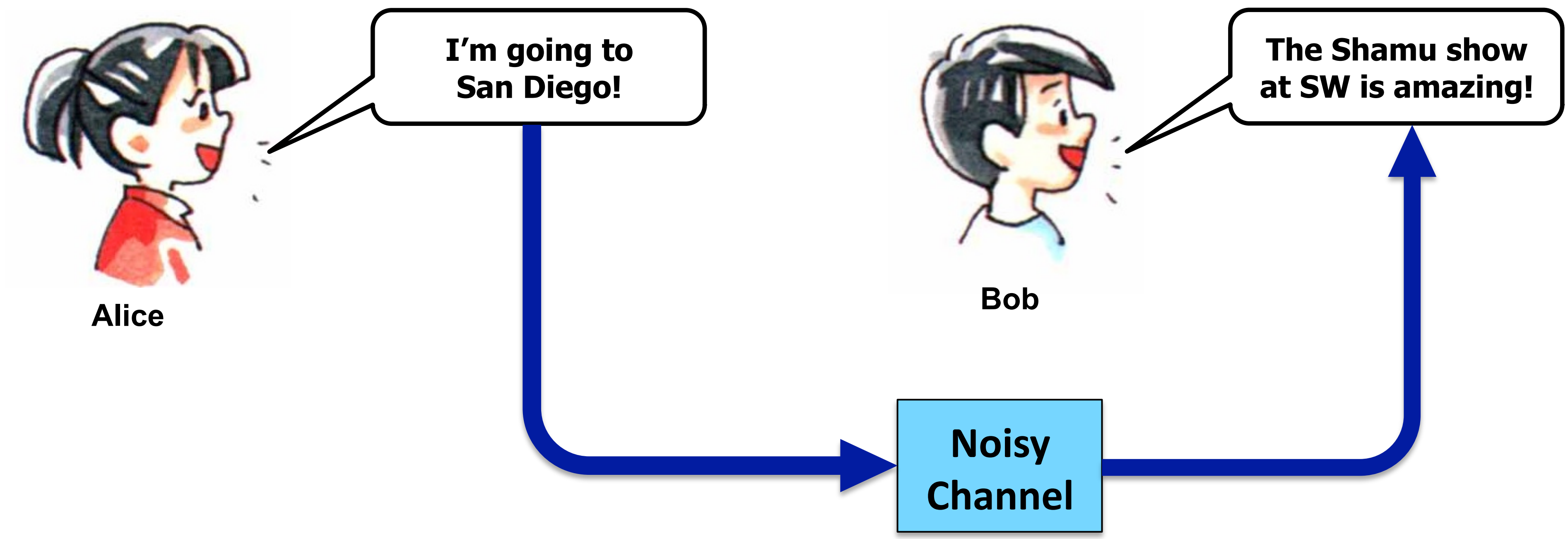
Total TE ≈ 0.025

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.

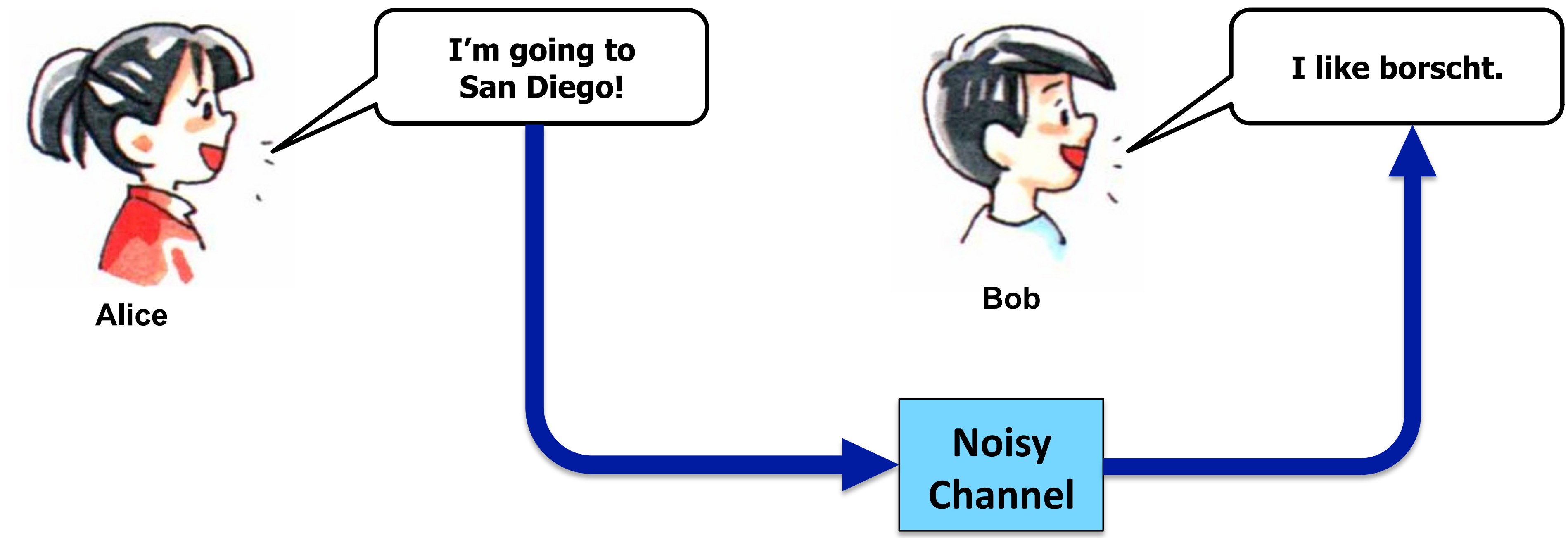


- Timing of Activity
- **Content Dynamics**
- Estimation of entropic measures (from limited data)

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:

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

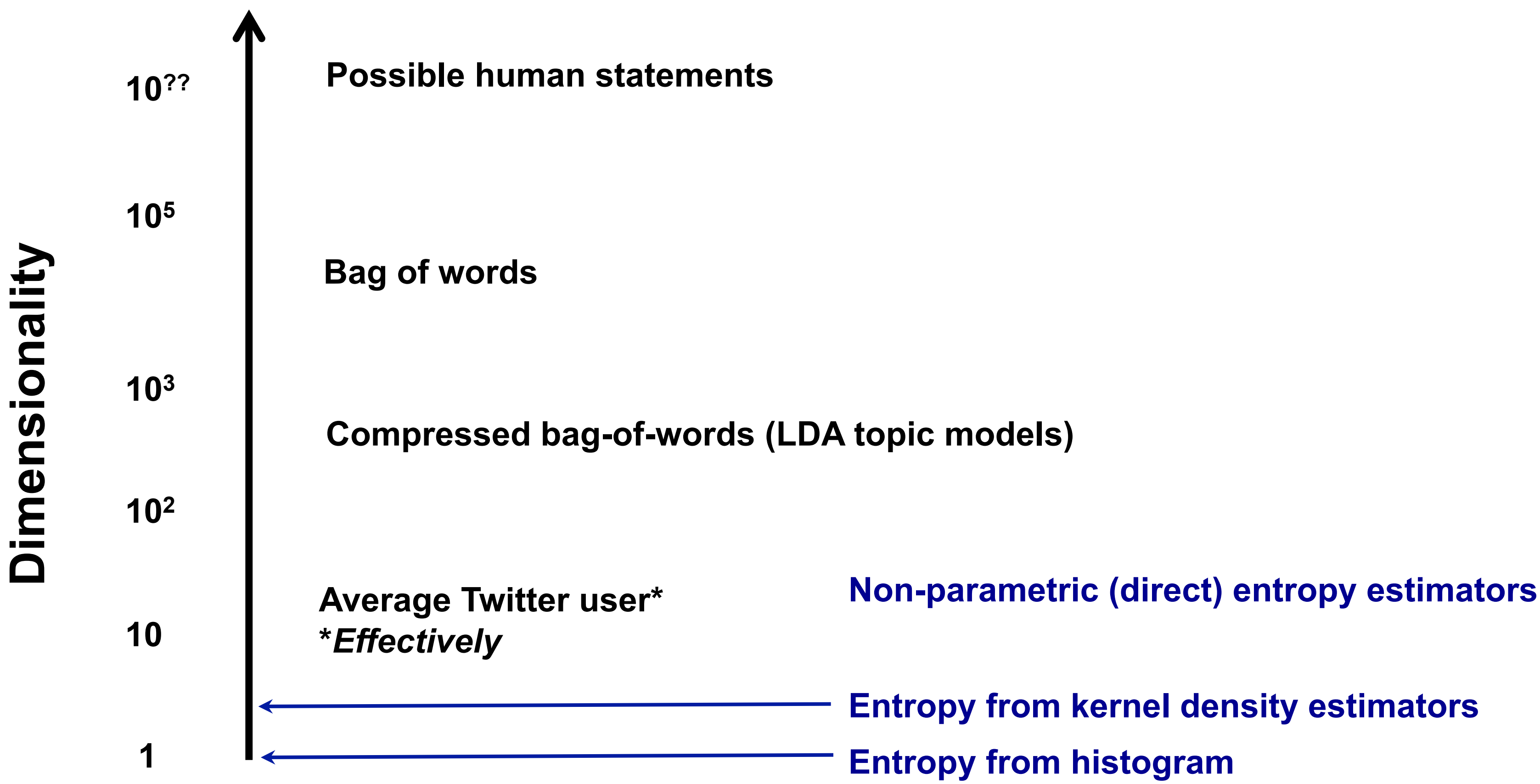
Sum over all possible statements!

- Includes such hard to quantify probabilities as:

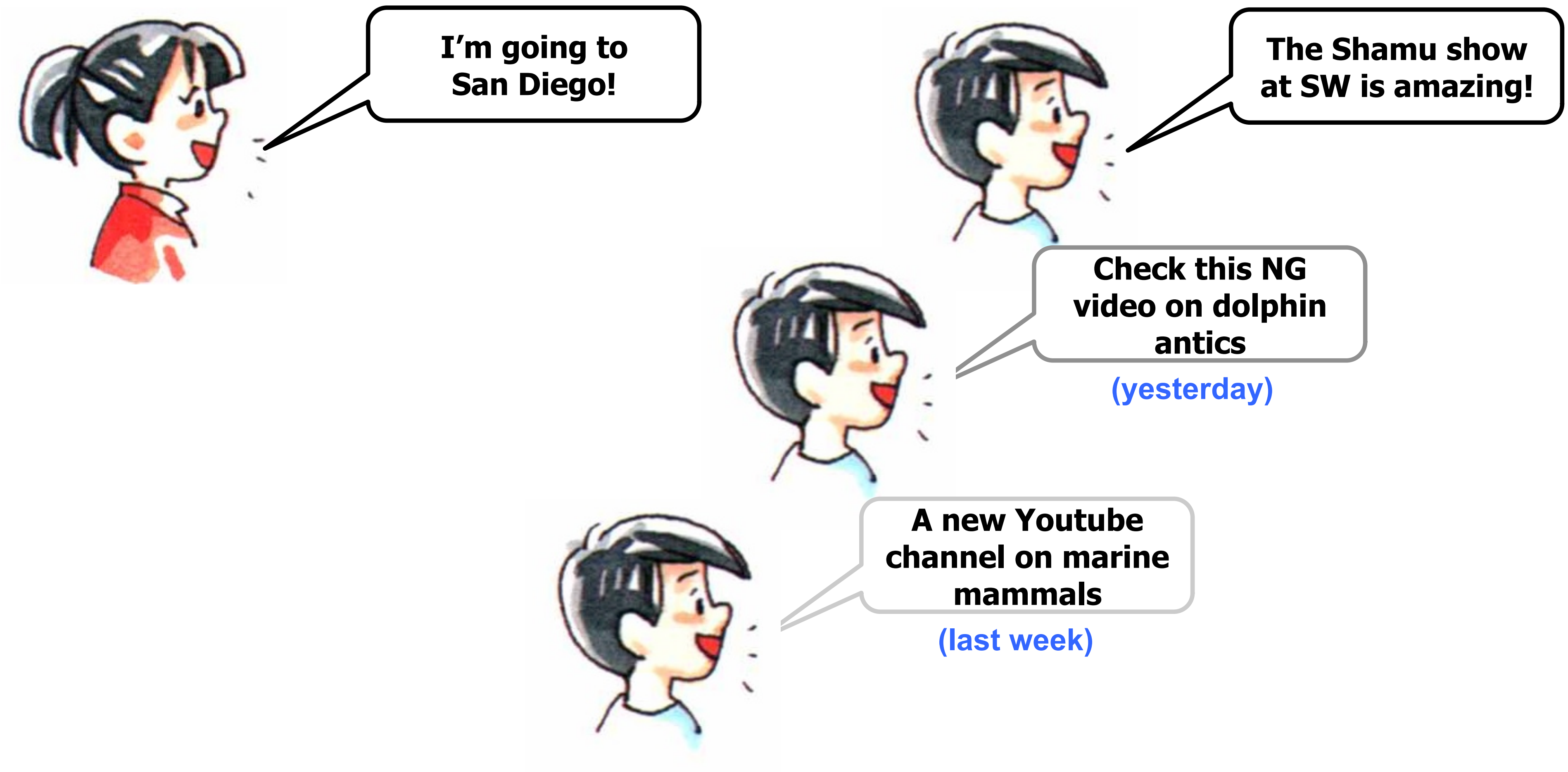
Pr(Alice says **"I'm going to San Diego**", then Bob says "**I like borscht**"**)**

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

You're so 10 dimensional

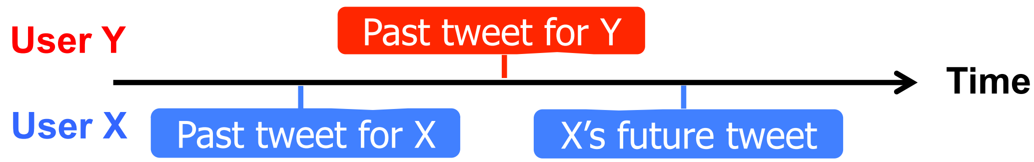


Information in human speech

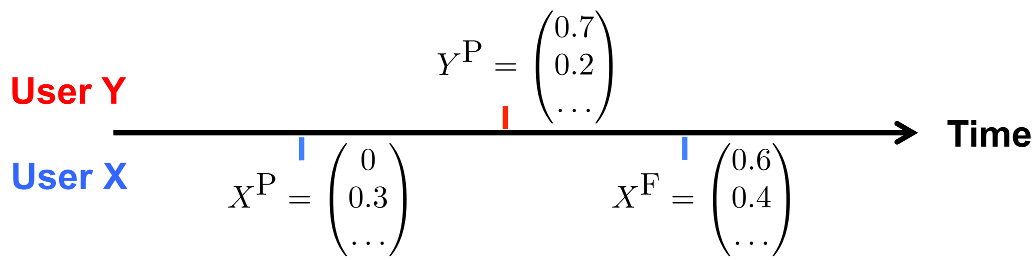


T.E. for Content Dynamics

- N samples of tweet exchanges



- Convert to an abstract representation

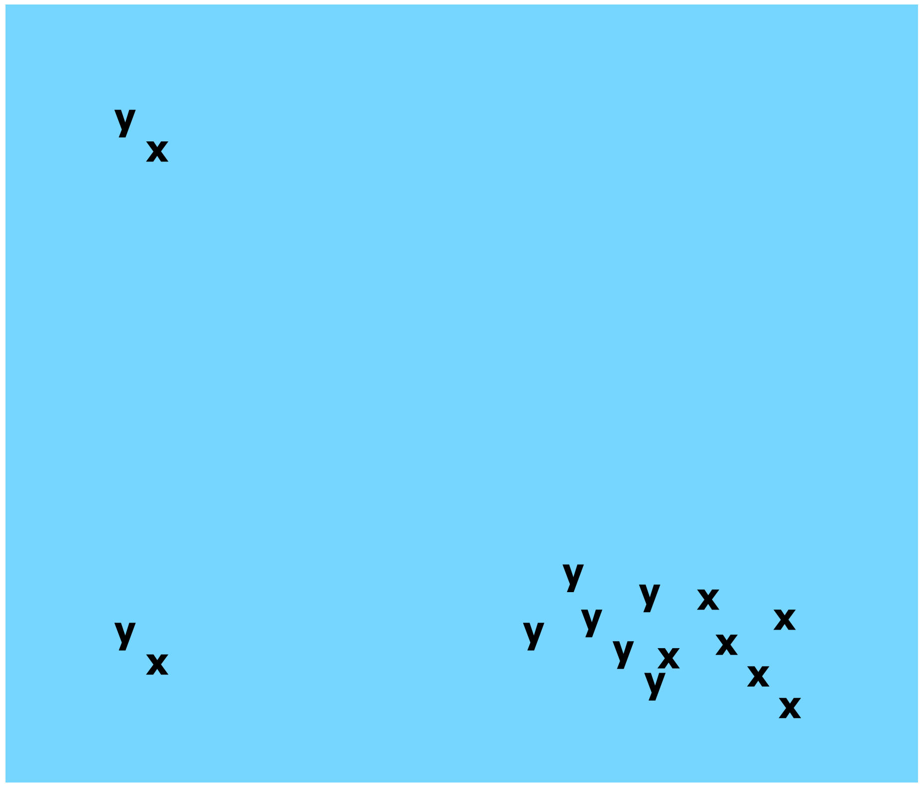


- Estimate transfer entropy: measure of Y's predictivity of X

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

Predictability in Content Space

Tweets about the 2012 election



Tweets about taxes

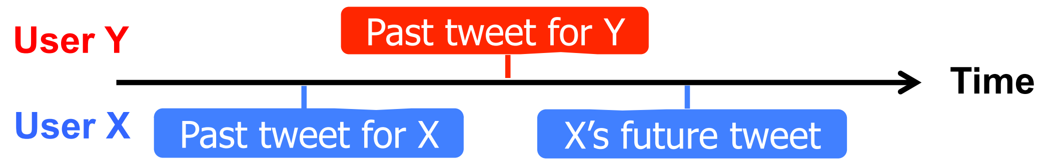
Tweets about health care reform

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

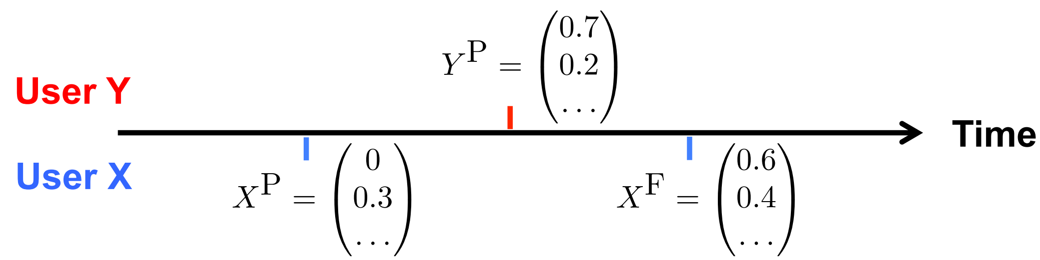
Low Transfer Entropy – x is already predictable

T.E. for Content Dynamics

- 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

HOLY FLYING COWS FROM SPACE WHY DID THIS SONG DO BAD IF IT'S SO INCREDIBLE.

$$\longrightarrow \begin{pmatrix} 0.01 \\ 0.32 \\ 0.61 \\ 0.04 \\ \dots \end{pmatrix}$$

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

Estimate transfer entropy

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

~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

Topic Modeling

- A *bag-of-words* representation for text

- A *document* is a sequence of *N* words denote by

$$\mathbf{d} = (w_1, w_2, \ldots, w_N)$$

- A *corpus* is a collection of M documents denoted by

$$D = \{\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_M\}$$

Latent Dirichlet Allocation

- Latent Dirichlet allocation (LDA) is a generative probabilistic model of a document corpus.
- Generative process for each document **d** in a corpus *D*:

1. Choose $N \sim \text{Poisson}(\xi)$ – number of words in **d**

2. Choose $\theta \sim \text{Dir}(\alpha)$ – the weights of different topics in **d**

3. For each of the N words w_n

   (a) Choose a topic $z_n \sim \text{Multinomial}(\theta)$

   (b) Choose a word w_n from $p(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic z_n

4. Inference and Learning

   (a) Topics and associated word probabilities

   (b) Topic mixture of each document

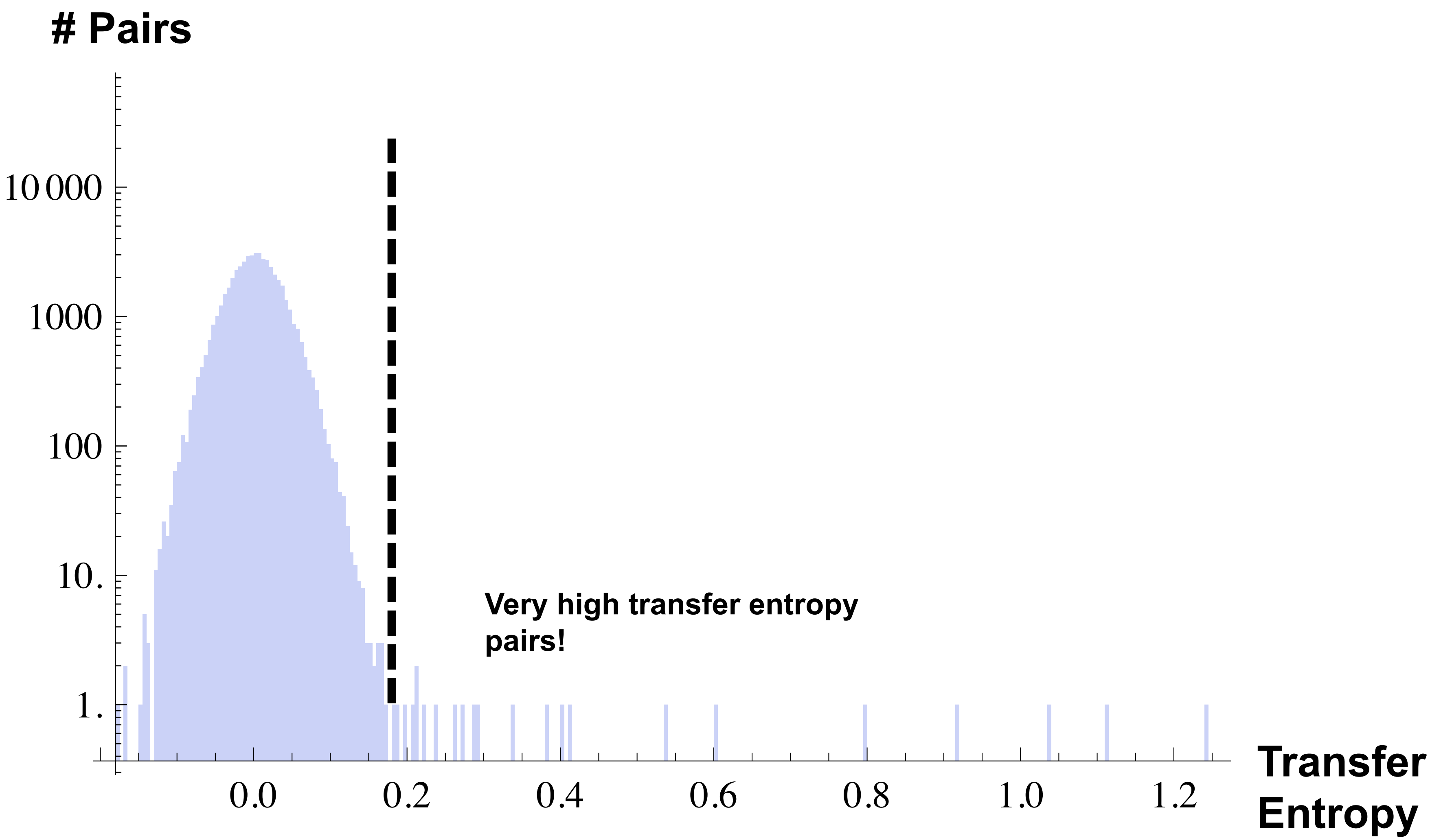
Twitter Study

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

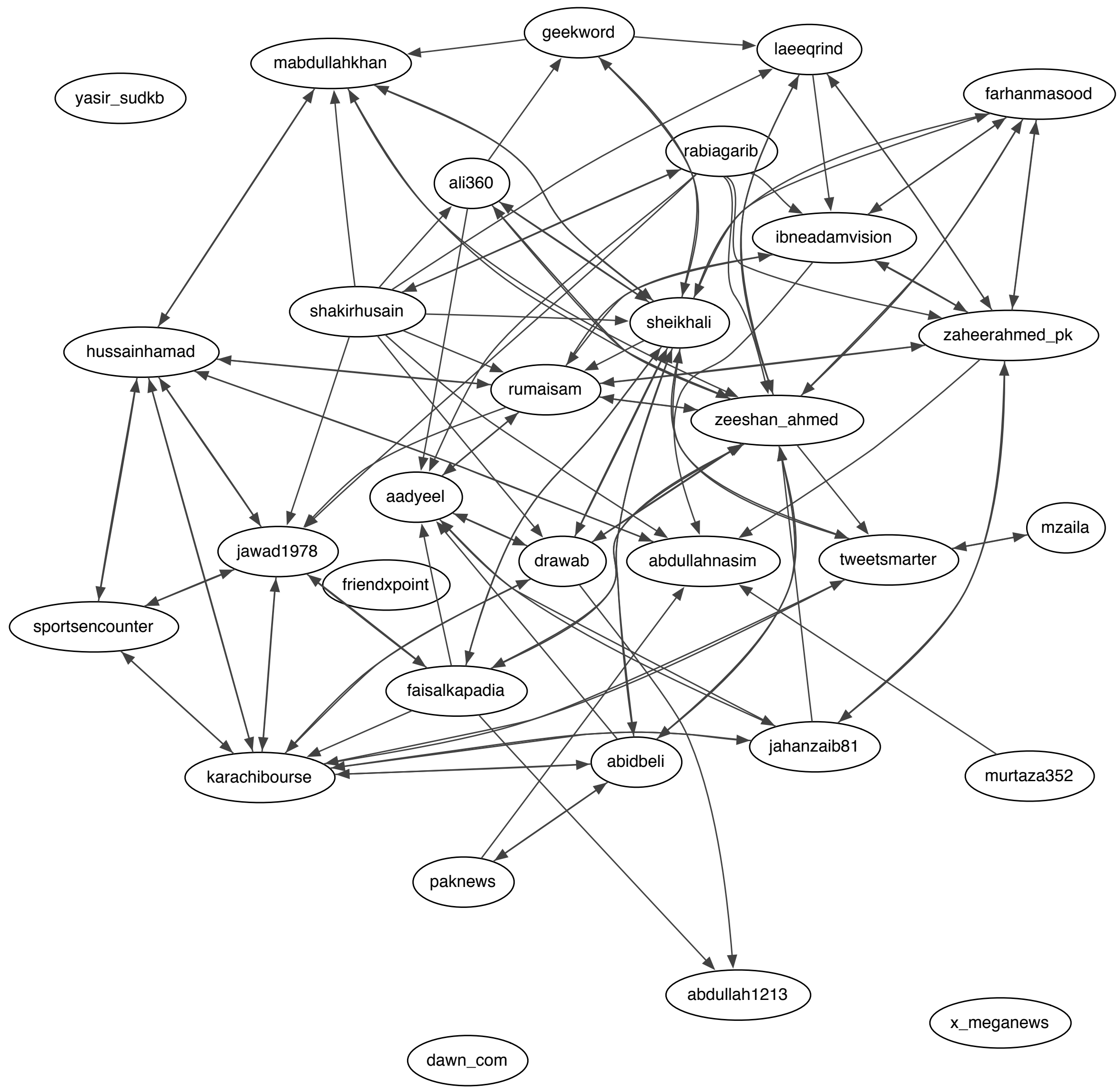
- PREPROCESSING:
  - **No RTs**
  - [a-zA-Z] only, lowercased
  - No punctuation
  - No stop words

- Calculate transfer entropy for all ordered pairs of users

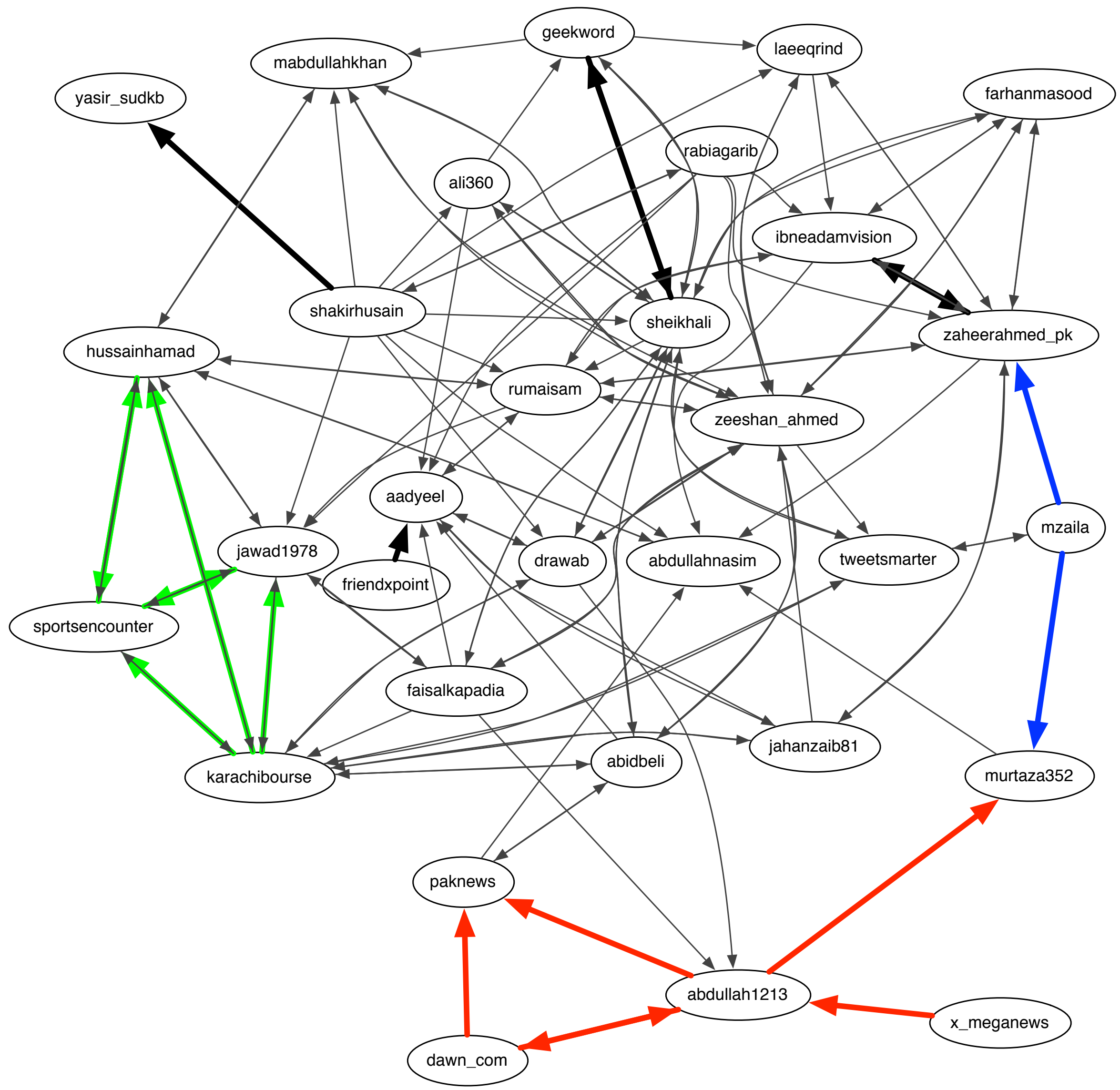
Histogram of transfer entropy



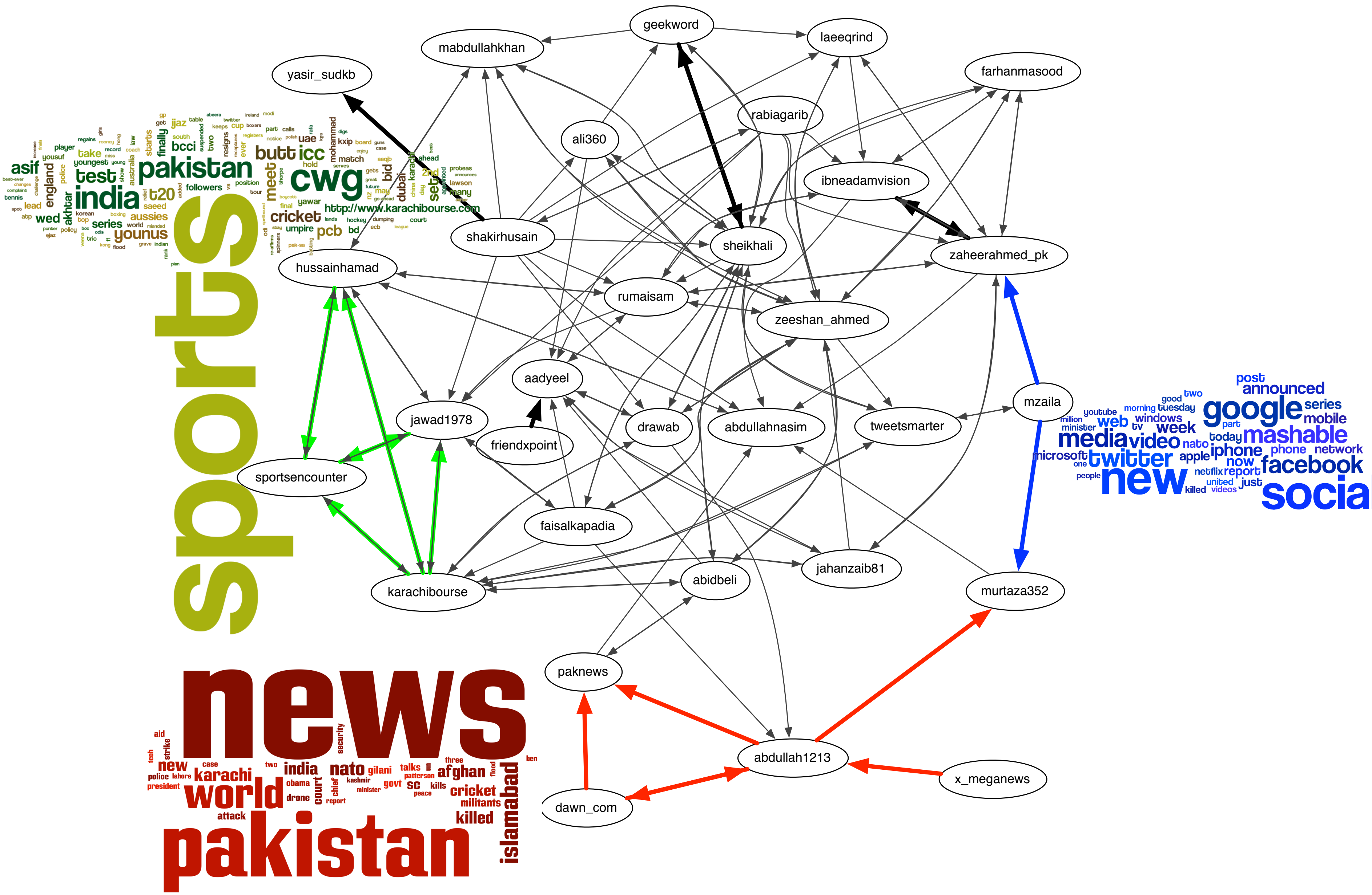
The “Friend” Network

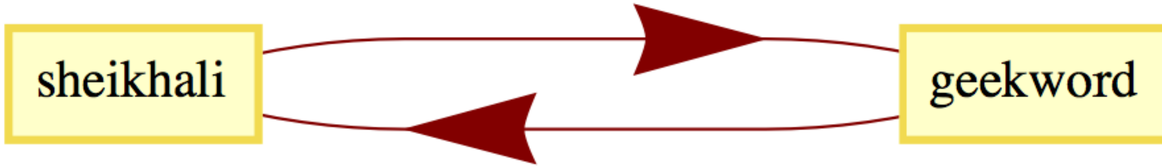


The Hidden Network (based on activity)



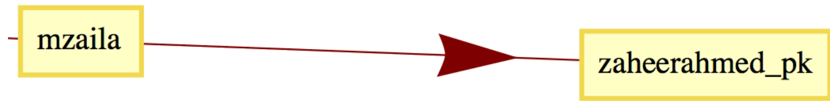
The Hidden Network (based on activity)





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

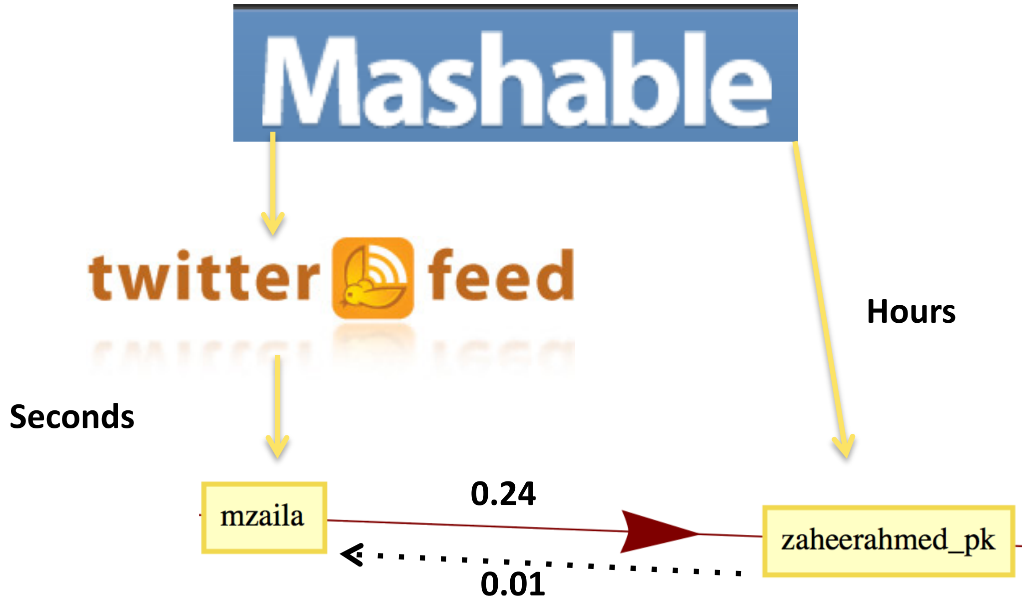
geekword: #Skype for #Windows gets deep rooted #Facebook Integration http://bit.ly/cb7UOj #SocialNetwork
sheikhali: #Skype for #Windows gets deep rooted #Facebook Integration http://bit.ly/cb7UOj #SocialNetwork
sheikhali: @l3v5y nice one
geekword: #Windows Phone 7 to get copy/paste feature in early 2011 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: ISLAMABAD, Pakistan, Oct. 14 (UPI) – Deforestation ... http://bit.ly/92afA0 #pkfloods #Pakistan
	mza	Google Shares Jump 7% on Impressive Earnings: Google

Social influence

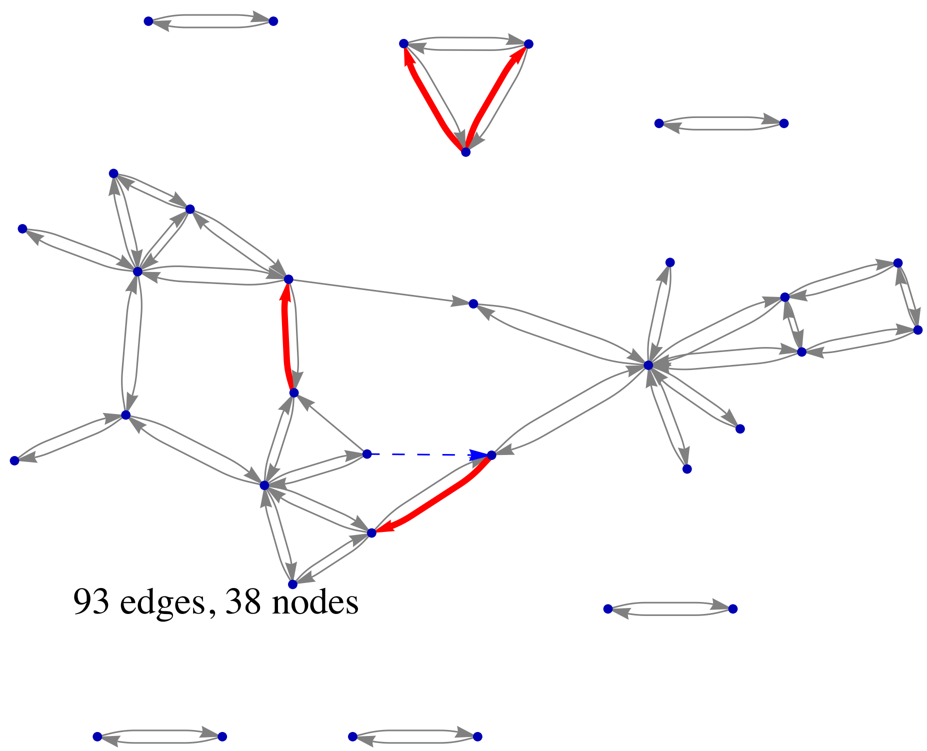
Previous examples were *predictable* but not *social*

- Can we use mentions to check if we capture social behavior?
- Mentions != Social

aya_bieber3: @justinbieber africa but not israel :(
aya_bieber3: @justinbieber i'm excited to see this video ♥ i love u
aya_bieber3: @justinbieber notice ur amazing isralis fans? (: ♥
aya_bieber3: @justinbieber i just want u to notice me or to ur fans in israel! but.. i guess u'll never do it :(
aya_bieber3: @justinbieber haha we have the same number of followers !! ♥
aya_bieber3: @justinbieber I will never say never until ull tweet me !!!
aya_bieber3: @justinbieber I will never say never until ull tweet me !! ♥
aya_bieber3: @justinbieber we have the same number of followers haha
aya_bieber3: @justinbieber I love uuuu <3
aya_bieber3: @justinbieber heyy justin how r u? ((:
aya_bieber3: @justinbieber it's weird but all the times u noticed me (2 times haha not really notice) were when i didn't mean u to do that (: love uu ♥
aya_bieber3: @justinbieber u know i love u? (:

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

Reconstructing mention graph



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**

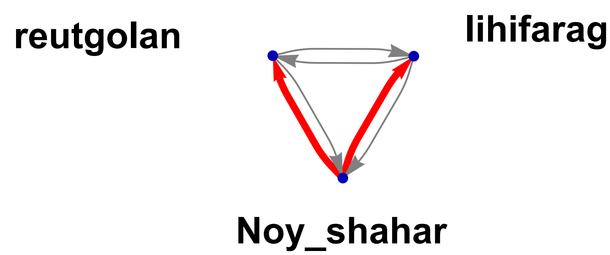
(w/ SE = 3.5%)

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



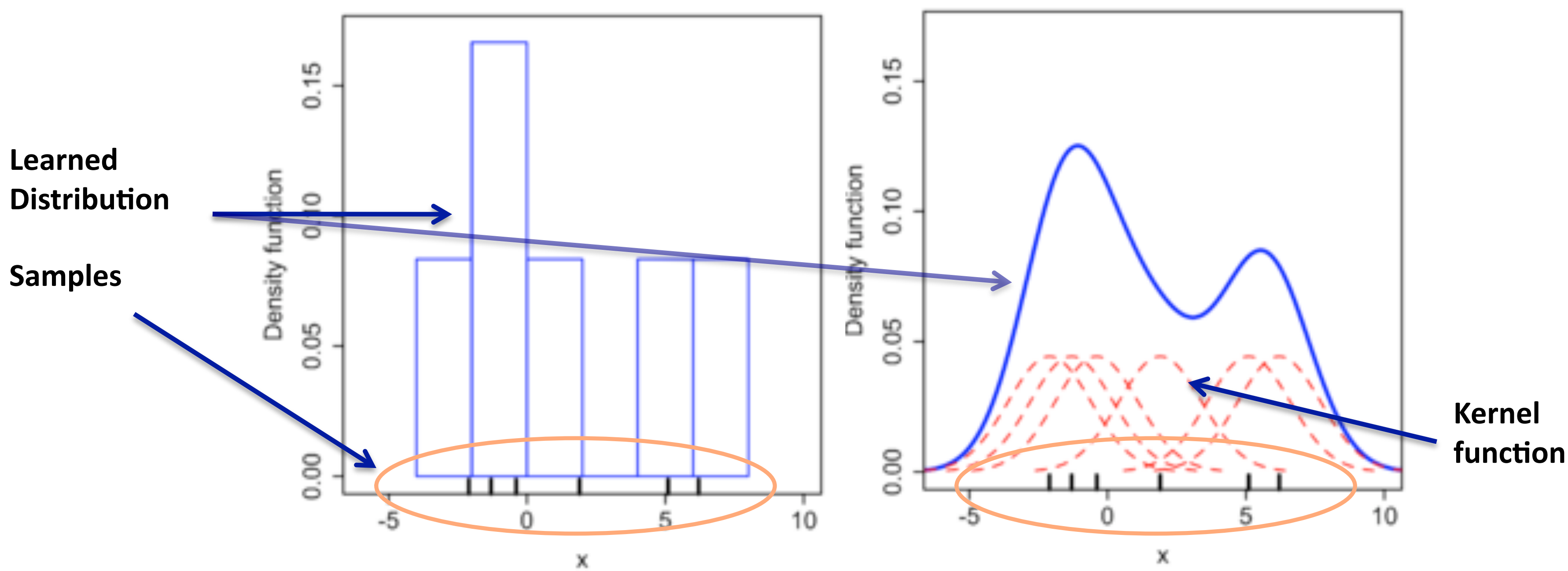
re	queremos unaa fotooooo deee @celeb1 y @celeb2
li	QUIERO UNA FOTO DE @celeb1 & @celeb2
no	@celeb2 nico .. please que la segunda imagen sera de vos con @celeb1
re	duele tanto decir ALGO ?
li	@celeb2 nico porfi saca una foto con emi :(
re	@No [Hebrew characters]
no	@Li @Re [Hebrew characters]
no	@re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

- Timing of Activity
- Content Dynamics
- **Estimation of entropic measures (from limited data)**

Problem

- We need probability distributions, usually we only have samples

$$H(x) = -\sum_{x} p(x) \log p(x)$$



Estimate entropies from samples?

Uncertainty about X

Uncertainty about X, if you know Y's behavior

$$\begin{aligned} TE_{Y \rightarrow X} &= H(X^{Future}|X^{Past}) - H(X^{Future}|Y^{Past}, X^{Past}) \\ &= CMI(X^{Future} : Y^{Past}|X^{Past}) \end{aligned}$$

Or, a conditional mutual information

Entropy is a functional of probability distribution, so, in principle, we have to first estimate:

$$p(X^{\mathrm{P}}, Y^{\mathrm{P}}, X^{\mathrm{F}})$$

Estimate entropies from samples?

Uncertainty about X

Uncertainty about X, if you know Y's behavior

$$
\begin{aligned}
TE_{Y \to X} &= H(X^{Future}|X^{Past}) - H(X^{Future}|Y^{Past}, X^{Past}) \\
&= CMI(X^{Future} : Y^{Past}|X^{Past})
\end{aligned}
$$

Or, a mutual information

But there's a better way:

Estimating Mutual Information

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger
John-von-Neumann Institute for Computing, Forschungszentrum Jülich, D-52425 Jülich, Germany
(Dated: February 2, 2008)

We present two classes of improved estimators for mutual information $M(X,Y)$, from samples of random points distributed according to some joint probability density $\mu(x,y)$. In contrast to conventional estimators based on binnings, they are based on entropy estimates from k-nearest neighbour distances. This means that they are data efficient (with $k=1$ we resolve structures down to the smallest possible scales), adaptive (the resolution is higher where data are more numerous), and have minimal bias. Indeed, the bias of the underlying entropy estimates is mainly due to non-

Intro to bin-less entropy estimator

One way to write entropy:

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

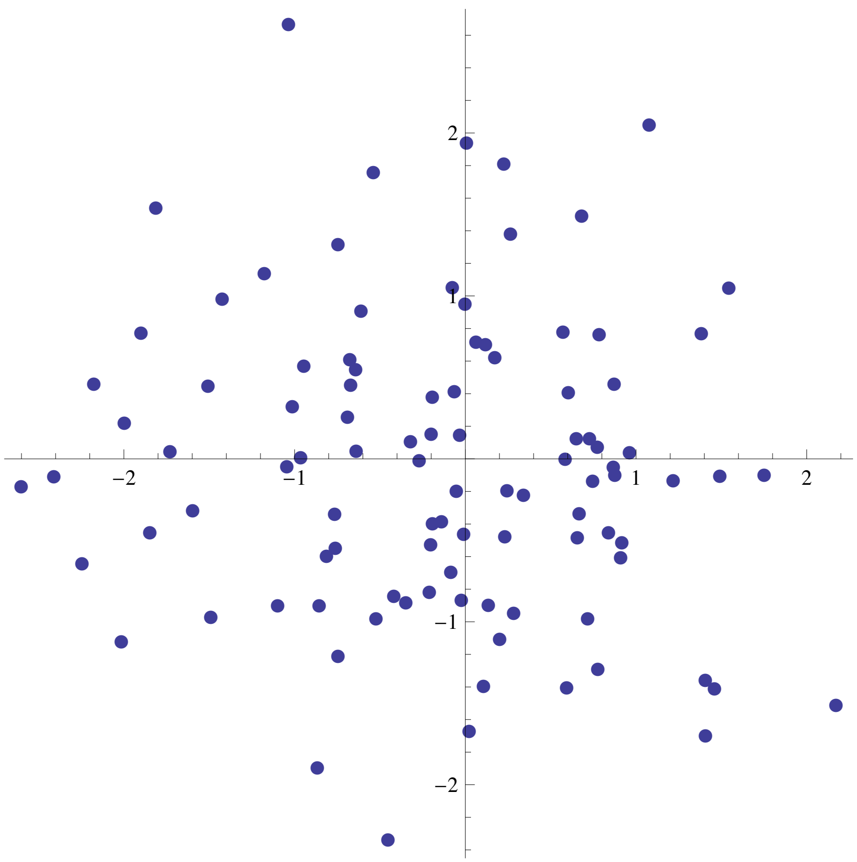
Given some samples $x_i \sim p(x)$,

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

But there's a problem, the whole point is we don't know p(x)

Intro to bin-less entropy estimator

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



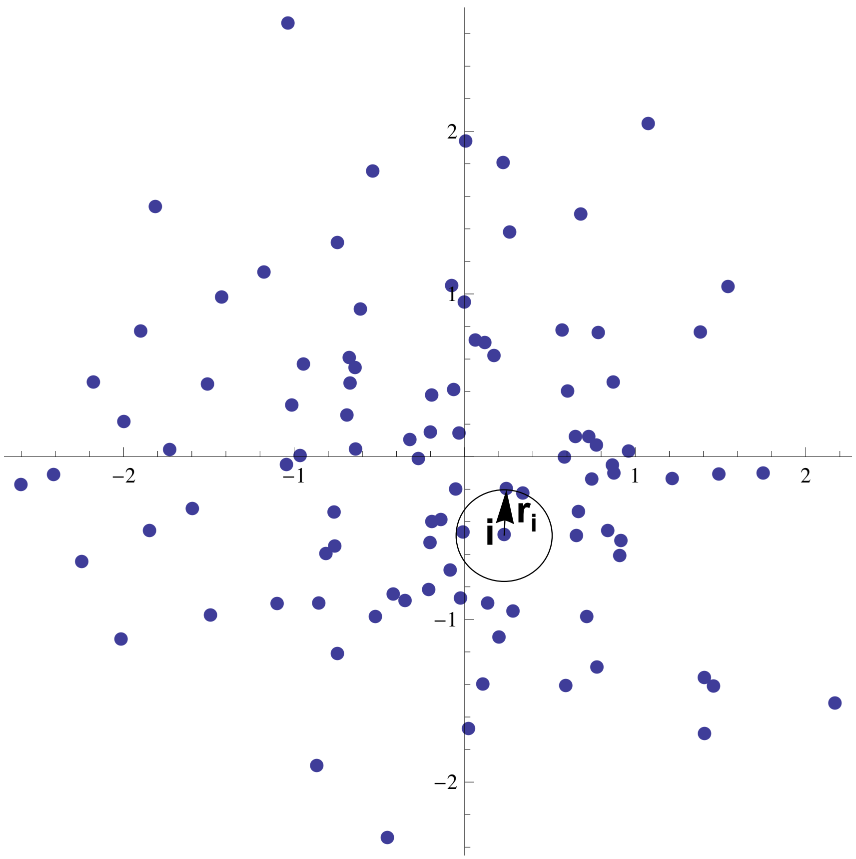
Instead, we'll estimate the density p(x) at each point x_i

$$\hat{p}(x_i)$$

Intro to bin-less entropy estimator

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

$$\propto \frac{d}{N}\sum_i \log r_i$$



Instead, we'll estimate the density p(x) at each point x_i

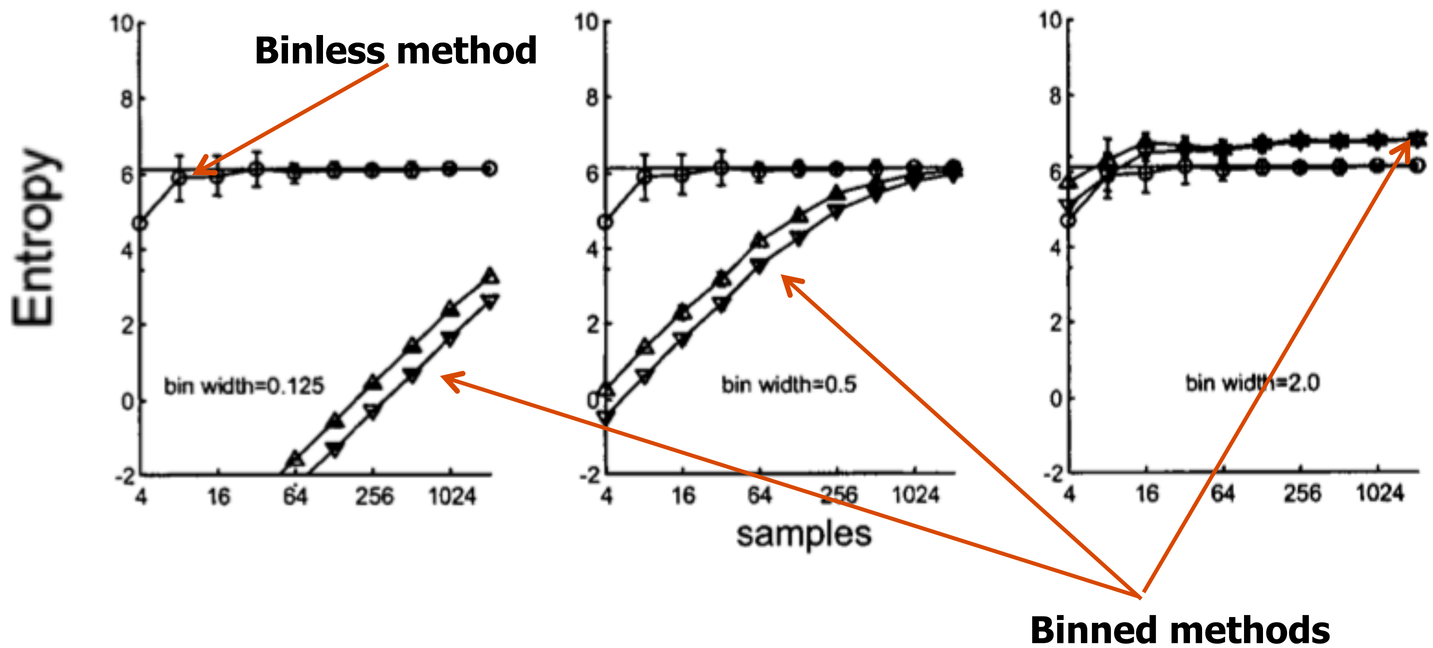
$$\hat{p}(x_i) = \frac{\%\text{ points in ball } i}{\text{Volume of ball } i}$$

$$k = 3$$

$$\hat{p}(x_i) \approx \frac{3/N}{\pi r_i^2}$$

Advantage of bin-less estimator

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



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

But for topic models?

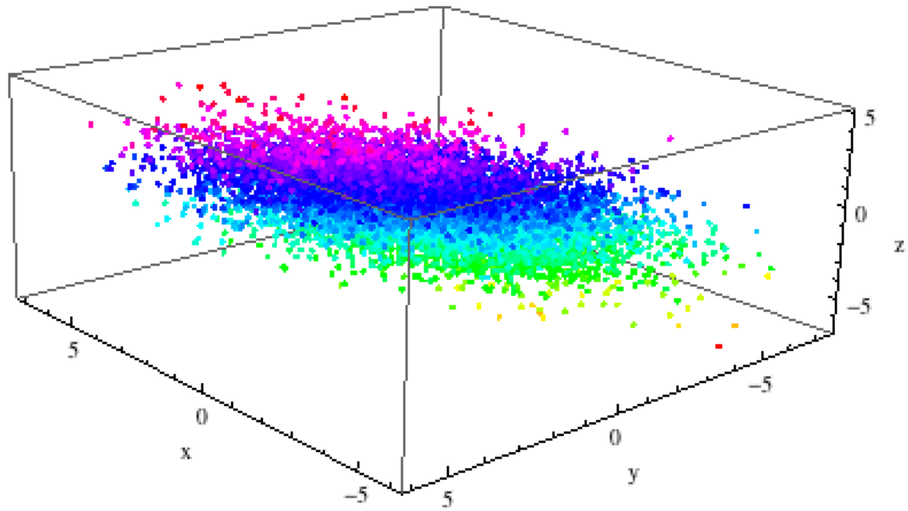
- Nice trick in a few dimensions, but if we pick a topic model with 125 topics,

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

- 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)
- If not, does it matter that we wrote it that way?
  - (answer: no! The estimator relies on distances only)

Example

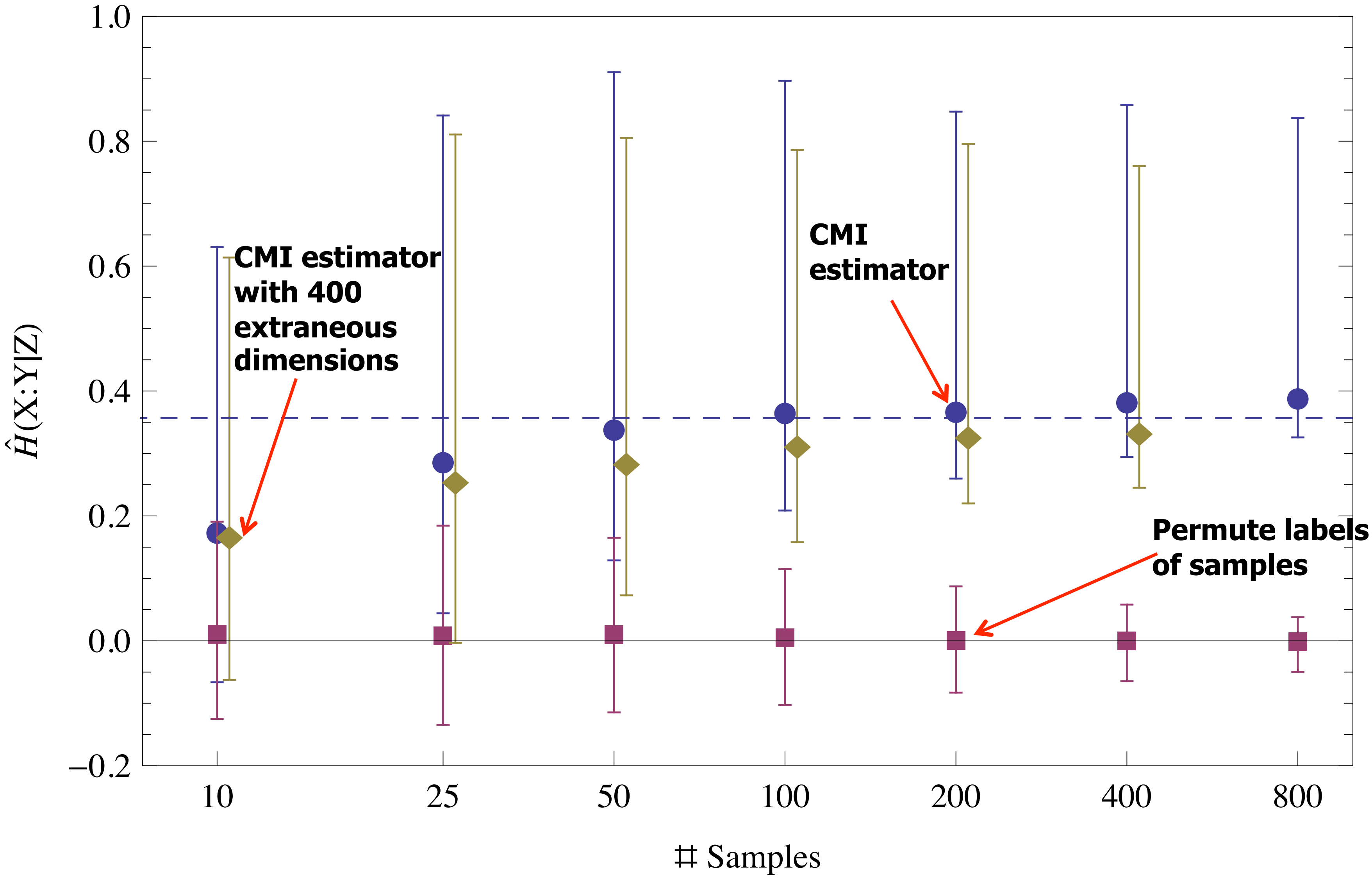


$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 4 & 3 & 1 \\ 3 & 4 & 1 \\ 1 & 1 & 2 \end{pmatrix} \right)$$

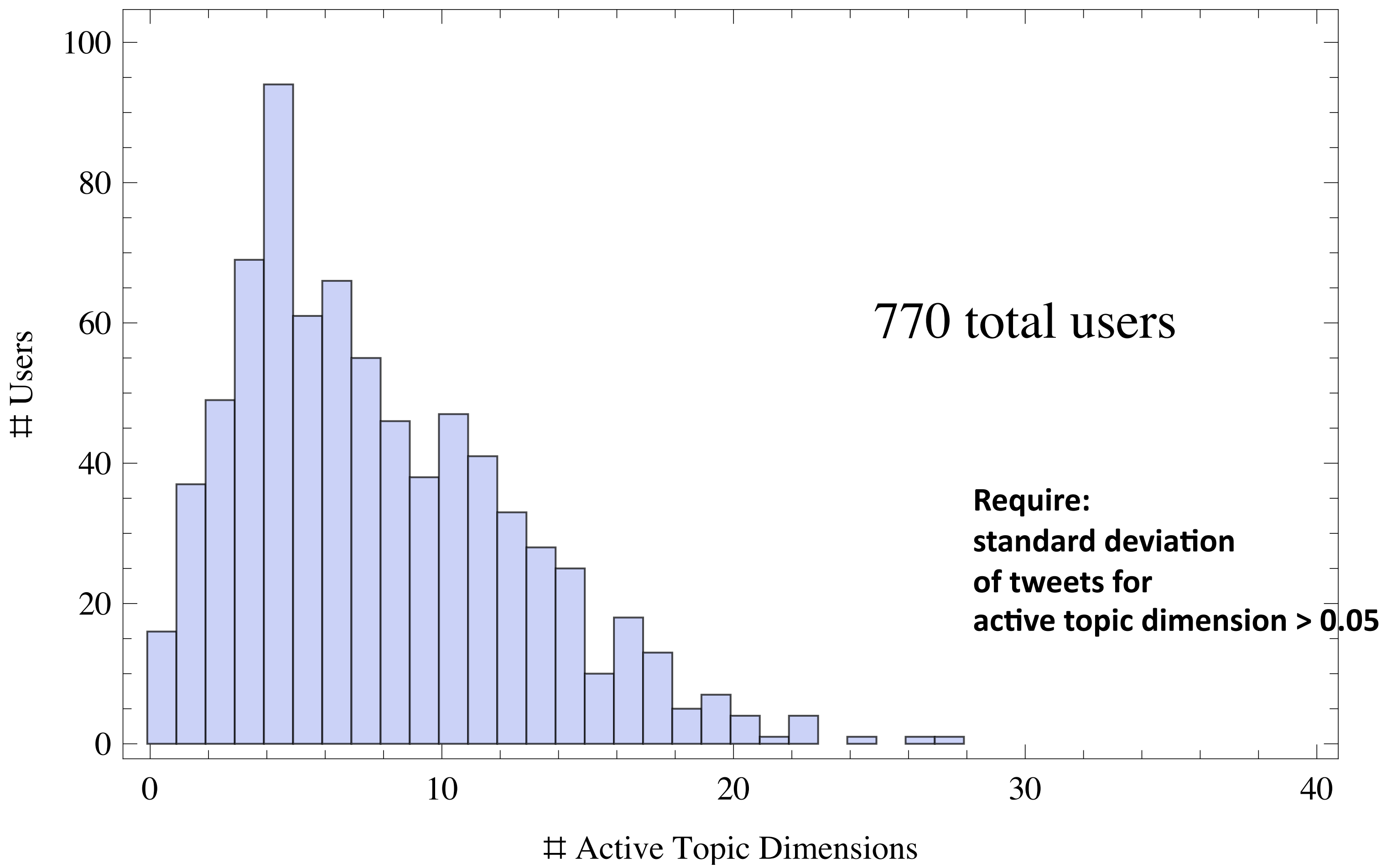
$$H(X : Y | Z) = 0.357$$

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

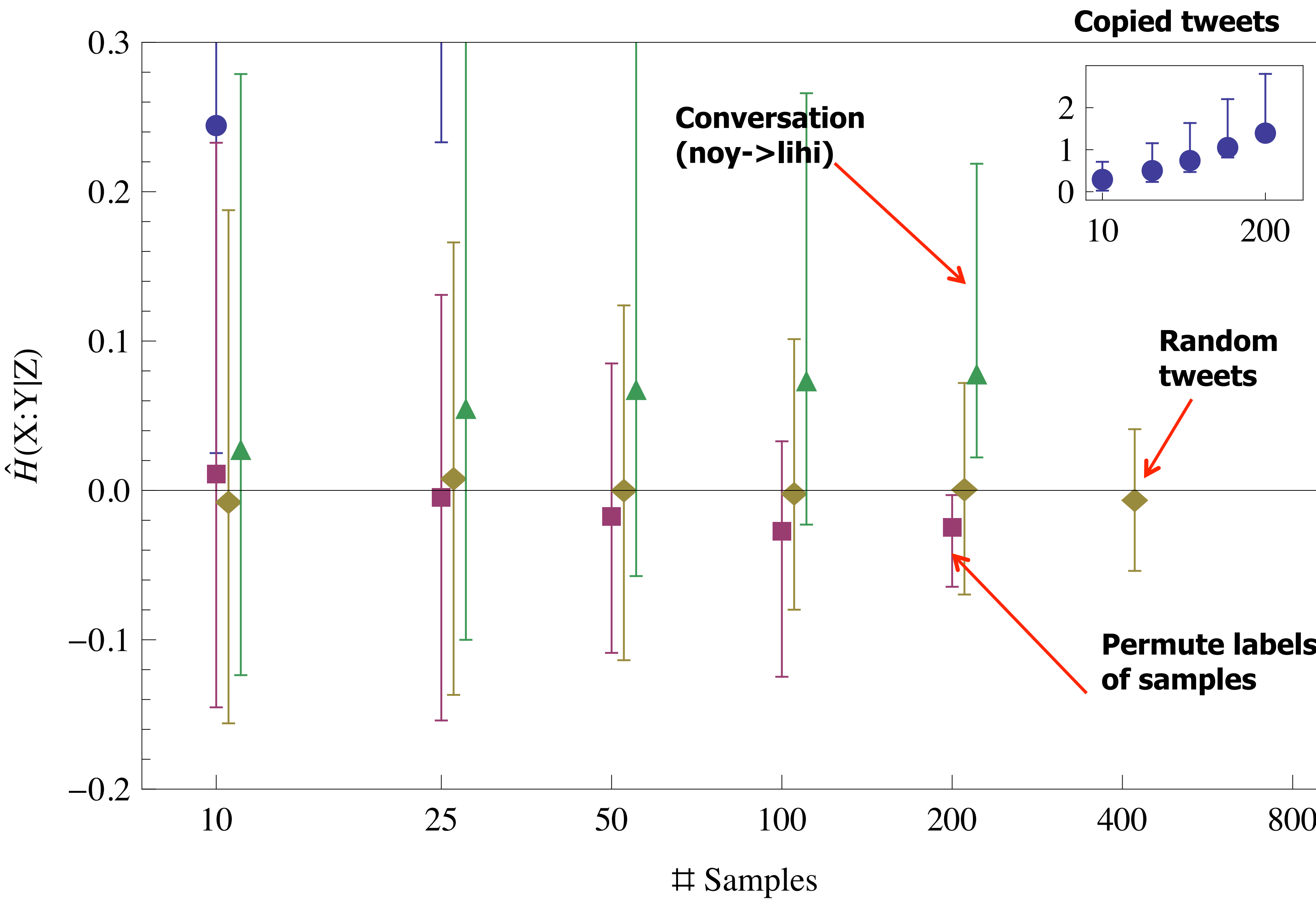
Convergence of estimators



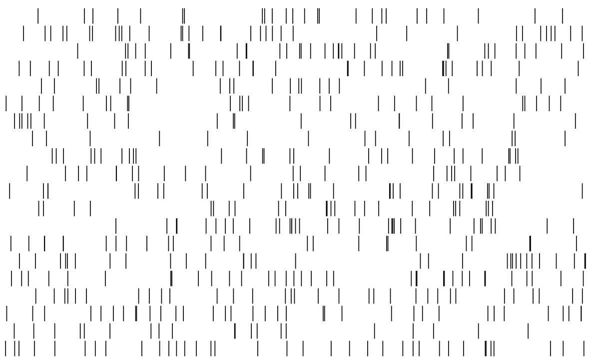
Number of active topics per user

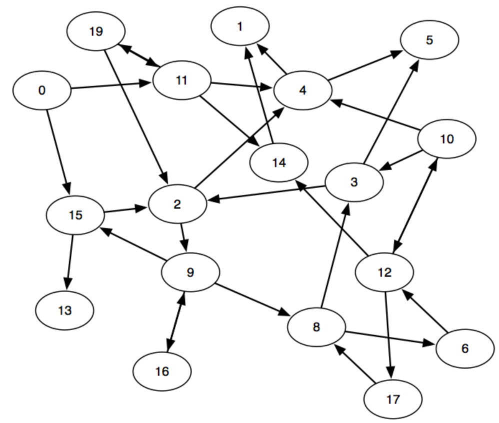


Convergence for some real-world data



Summary





tabankhamosh: PEPCO plants in the seraiki belt belch un-scrubbed pollutants into the air, sicken the citizens & dump sludge in the water table. #taraqqi

enggandy: NEW VERSION OF TWITTER IS HERE ...

shahidsaeed: @[redacted] ISI officers will be sent to PEPCO, WAPDA, NEPRA, PTA, PR, PIA, PSM, NBP, etc and fix everything. ISI zindabad

tabankhamosh: Revenge of the Seraikis = The Punjabi strain of the Taliban. #oy #vey

enggandy: @[redacted] YEAH I THINK SO ... YOU GOT IT ??? SPLIT SCREEN VERSION ?

shahidsaeed: @[redacted] Seraikis meant for cultivating fields so Punjab speaking South-Punjabi people fight Jihad in Kashmir,Chechya,Phillipines,etc

fzzzkhan: u got it? :O RT @[redacted] NEW VERSION OF TWITTER IS HERE ..

fzzzkhan: @[redacted] no :(m still waiting for it

Transfer entropy:

- Recover *predictive links* from user activity
- Grounded in information theory, can work for arbitrary signals
  - Timing of activity
  - Generated content

Ongoing and Future Work:

- Different representation of content (e.g., stylistic features)
- Better TE estimators

Thank you. Questions?

Pre-prints bit.ly/Qc8s84
bit.ly/pgYtJP

Code: bit.ly/SmuOrr

Contact: {galstyan,gregv}@isi.edu