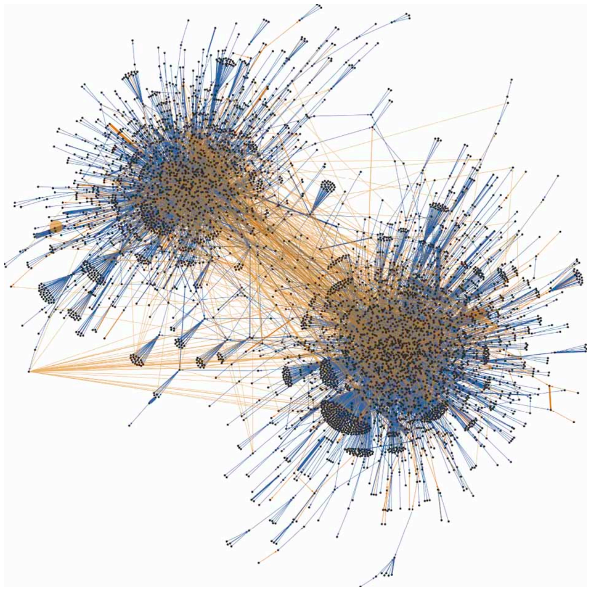
Quantifying the Impact of Rare Causes

Samantha Kleinberg
Stevens Institute of Technology

Rare events are common



Twitter: 8 TB/day

Financial markets

ICU: 5 sec measurements

Why causality?

- Usual approach: data mining
  - *Identify* rare events
  - E.g. credit card fraud, network intrusions
- But for action…
  - How do events *affect* system?
  - E.g. medical treatment, public policy

Why not use current methods?

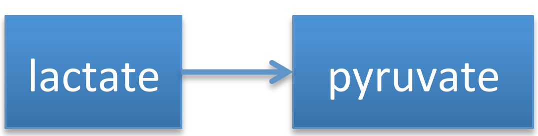
- Rare event mining
  - No information about impact of events
  - False alarms
- Causal inference
  - Probabilistic: can't handle infrequent events
  - Granger: assesses whole time series
  - Complexity: efficiency is key with big data

What's needed?

- Way of assessing events that may only occur once or twice
- Ability to distinguish between rare event and unmeasured variables
- Assessment of statistical significance
- Method for assessing immediate impact vs. regime change

Linking type and token

- Type-level model gives expectation for each timepoint
- Difference between actual observation and prediction is what’s omitted – rare events, hidden variables



L↑P↑....L↓P↓...L↑P↓

Overview

- Infer “normal” model
  - Use huge volume of data
- Find how much is not explained by model
- Determine how explanatory rare event is
  - Comparing to average distinguishes between unmeasured and infrequent events

Normal model

- Causes of continuous-valued effects
  - Average difference cause makes to value, rather than probability of effect

$$c \leadsto^{\geq r, \leq s}_{\geq p} e > E[e]$$

$$\varepsilon_{avg}(c,e) = \frac{\sum_{x \in X \;\; c} E[e|c \wedge x] - E[e|\neg c \wedge x]}{|X \backslash c|}$$

- Remove timepoints immediately after (or around) effect when doing inference

Kleinberg IJCAI, 2011

How explanatory is model?

- For each instance of variable, compare value to that predicted by model

$$\frac{\sum_{e_t} E[e_t] - e_t}{\#e}$$

$$E[e_t] = \text{sum of } \varepsilon_{avg}(a, e)$$

A = set of causes instantiated for time t

e_t = actual value of e at time t

What is impact of rare event?

- How much of E is unexplained after rare event r (after accounting for known causes)?

$$\frac{\sum_{e_t}(E[e_t] - e_t|r)}{(\#e|r)}$$

- Compare to average unexplained value
  - Event with no impact will not differ significantly
  - Factors out unmeasured variables

Computational complexity

- $O(r)$, r= number of occurrences of rare event
- $O(n^3T)$ for finding normal model

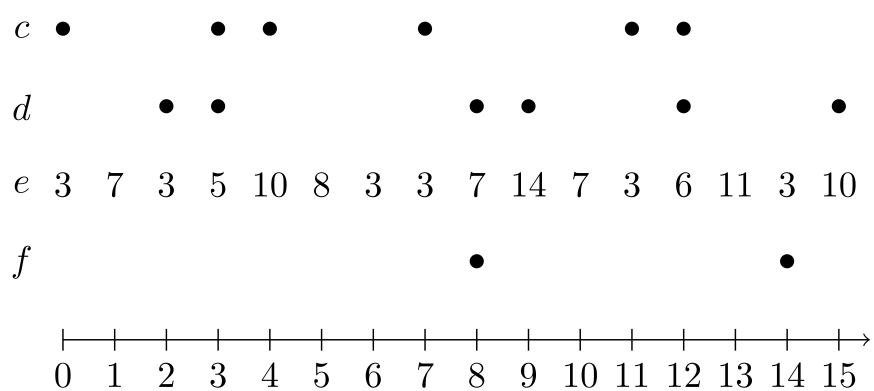
Example

- Normal model

$$c \rightsquigarrow e \qquad \varepsilon_{avg}(c,e) = 4$$

$$d \rightsquigarrow e \qquad \varepsilon_{avg}(d,e) = 3$$

- Observations

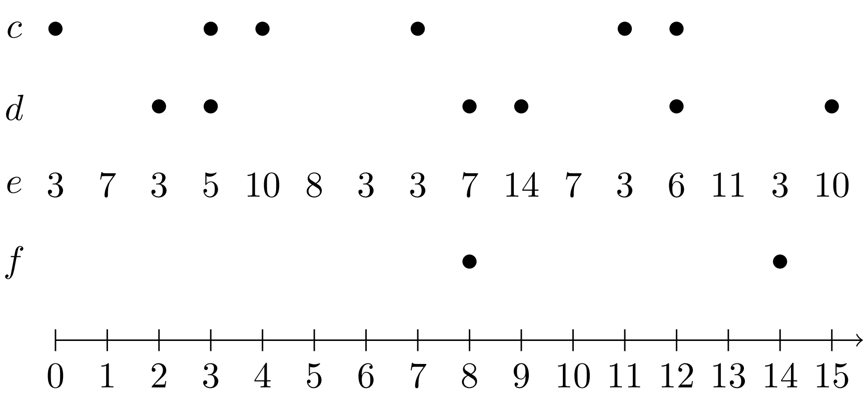


Example

- Normal model and observations

$$c \rightsquigarrow e \qquad \varepsilon_{avg}(c, e) = 4$$

$$d \rightsquigarrow e \qquad \varepsilon_{avg}(d, e) = 3$$



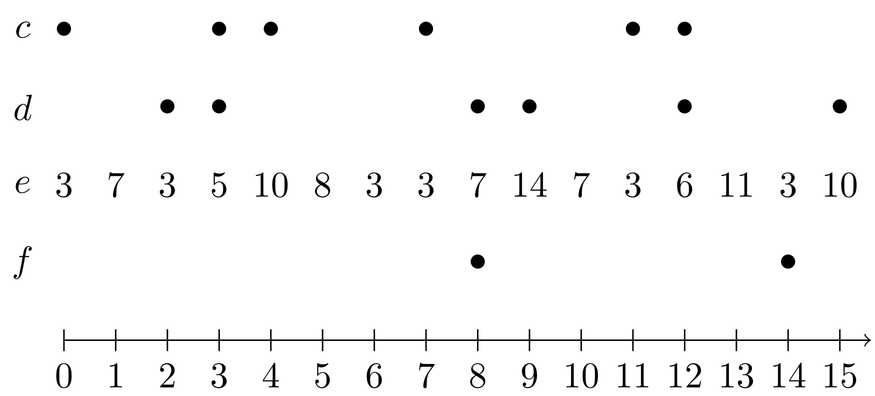
- Average unexplained value of e: 4

Example

- Normal model and observations

$$c \rightsquigarrow e \qquad \varepsilon_{avg}(c, e) = 4$$

$$d \rightsquigarrow e \qquad \varepsilon_{avg}(d, e) = 3$$



- Average unexplained value of e: 4
- Average unexplained value of e after f?

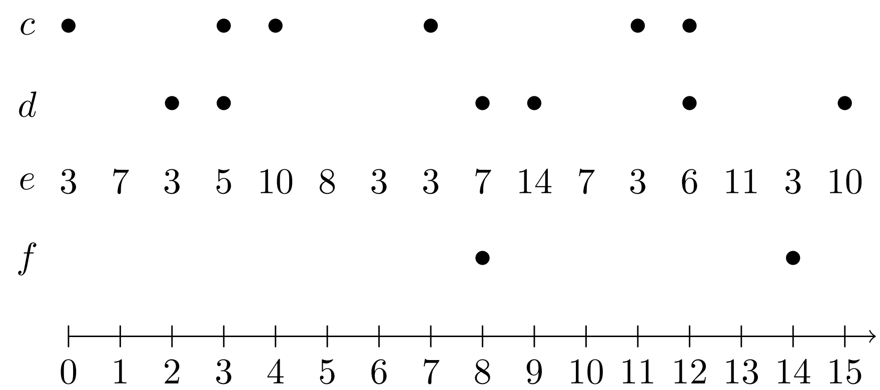
  11+10/2 = 10.5

Example

- Normal model and observations

$$c \rightsquigarrow e \qquad \varepsilon_{avg}(c,e) = 4$$

$$d \rightsquigarrow e \qquad \varepsilon_{avg}(d,e) = 3$$



- Average unexplained value of e: 4
- Average unexplained value of e after f?

  11+10/2 = 10.5

  p-value (from unpaired t-test): 0.0035

Simulated data

- Synthetic data based on Fama-French factor model [Fama & French 1992]

$$r_{i,t} = \sum_j \beta_{ij} f_{j,t} + \varepsilon_{i,t}$$

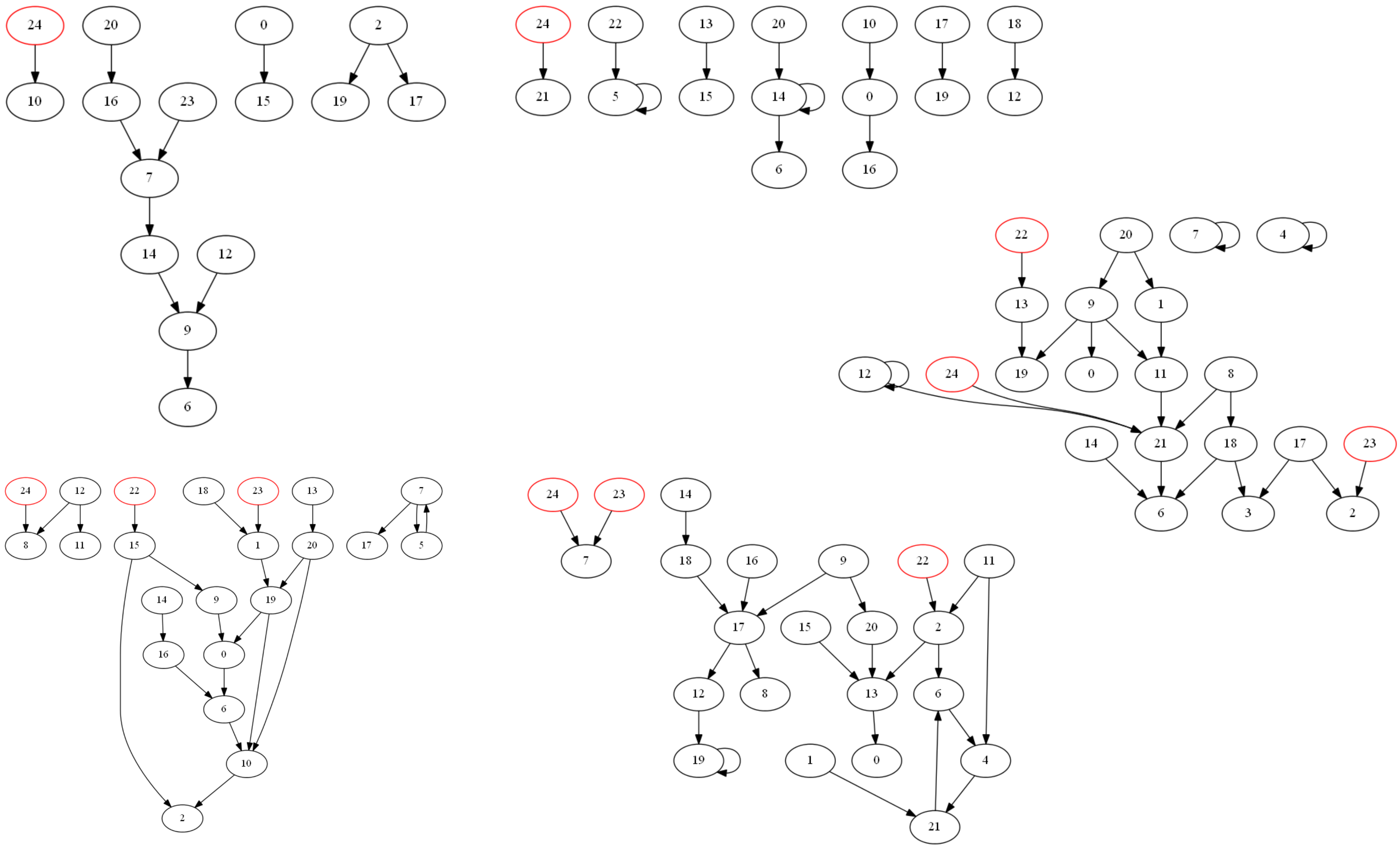
- Causality
  - Through epsilons (return of stock i at time t depends on return of stock j at time t-1)
  - Through constant term (return of stock i at time t increases by set amount if j is up at time t-1)

Experiments

- Simulated financial time series data
  - 5 structures [next slide]: 2 with 10 causal relationships, 3 with 20
  - 1-3 rare causes in each
  - 3 different probabilities for rare events (0.005, 0.0025, 0.0005)
  - 25 variables 4,000 time points

- 60 datasets (5 structures, 3 probabilities, 4 runs each)

Structures


24 20 0 2
10 16 23 15 19 17
7
14 12
9
6
24 22 13 20 10 17 18
21 5 15 14 0 19 12
6 16
22 20 7 4
13 9 1
12 24 19 0 11 8
14 21 18 17 23
6 3 2
24 12 22 18 23 13 7
8 11 15 1 20 17 5
14 9 19
16 0
6
10
2
24 23 14
7
18 16 9 22 11
17 15 20 2
12 8 13 6
19 1 0 4
21

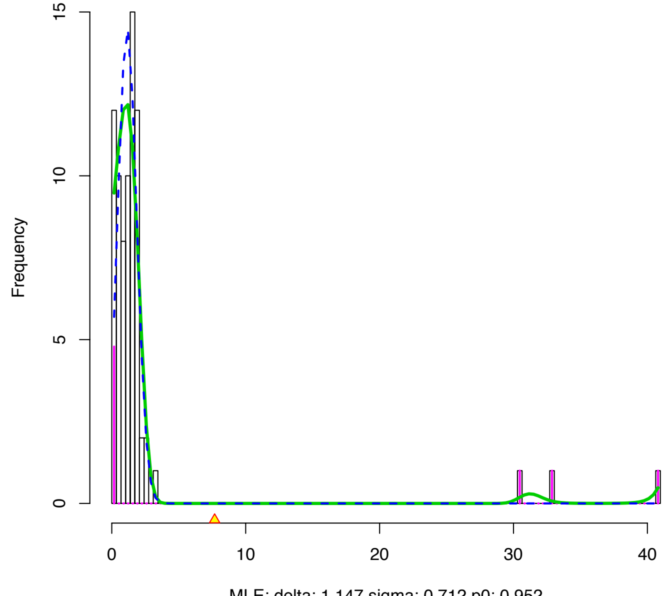
Results

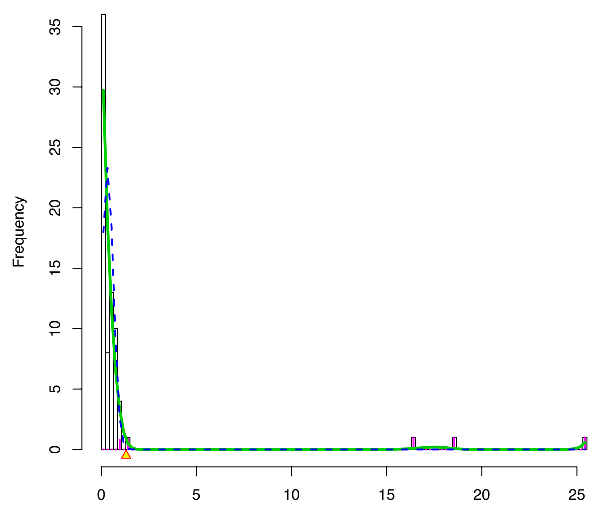
- 1-recall: proportion of sig. rare events discovered out of all embedded
- In 4K events, prob 0.0025 = Expectation of 2 occurrences.
- T-stat can't be calc with event that occurs once (this happened 7 times, didn't occur at all 5 others). This accounts for 12/16 FNS. Effective rate 9% when p=0.0005

P (rare event)	FDR	1-recall
0.005	0	2/44 (~5%)
0.0025	0	4/44 (~9%)
0.0005	0	16/44 (~36%)
TOTAL	0	22/132 (~17%)

Results

- 10,000 timepoints and two different time periods (120 datasets)
- Results
  - FDR: 2/262 (~.008%)
  - 1-recall: 4/264 (~.015%)
  - Normal model: FDR ~8%, 1-recall ~ 13%



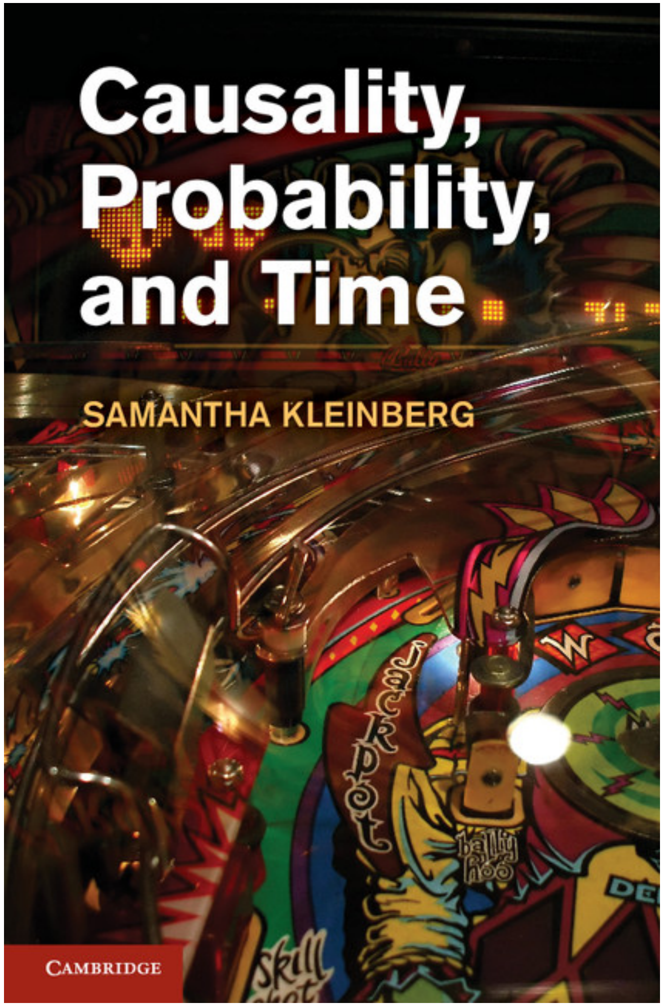


Conclusions

- Rare events are prevalent, important for big data, and need to be understood causally
- Feasible to infer impact with low FDR, even with errors in normal model

- Future work
  - Nonlinear relationships
  - Continuous + discrete variables
  - Latent variables

Also…



Published by Cambridge University Press

Out in November 2012!