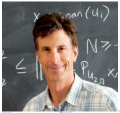
Some Mathematical Models to Turn Social Media into Knowledge

Xiaojin Zhu

University of Wisconsin–Madison, USA

NLP&CC 2013

Collaborators

Amy Bellmore	Aniruddha Bhargava	Kwang-Sung Jun	Robert Nowak	Jun-Ming Xu
				

Outline

1. Spatio-Temporal Signal Recovery (Poisson Generative Model)

2. Finding Chatty Users (Multi-Armed Bandit)

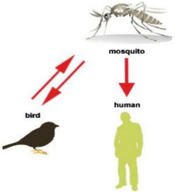
Outline

1. Spatio-Temporal Signal Recovery (Poisson Generative Model)

2. Finding Chatty Users (Multi-Armed Bandit)

Spatio-temporal Signal: When, Where, How Much

Public Health



"**100** dead robins found in **New York last Friday**"

Transportation Safety

"**16** deer got run over by cars in **Wisconsin last month** "

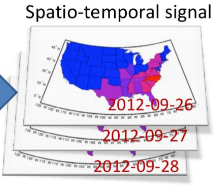
- Direct instrumental sensing is difficult and expensive

Social Media Users as Sensors

Text Pre-process

Time	Location
2012-09-26 17:35:23	Wisconsin US
2012-09-27 12:17:52	N/A
2012-09-27 08:28:12	(-98.24, 23.22)
...	

Socio-scope



- Not "hot trend" discovery: We know what event we want to monitor
- We are given a reliable text classifier for "hit"
- Our task: precisely estimating a spatiotemporal intensity function f_{st} of a pre-defined target phenomenon.

Challenges of Using Humans as Sensors

- Keyword doesn't always mean event
  - I was just told I look like dead crow.
  - Don't blame me if one day I treat you like a dead crow.
- Human sensors aren't under our control
- Location stamps may be erroneous or missing, e.g., in Twitter
  - 3% have GPS coordinates: (-98.24, 23.22)
  - 47% have valid user profile location: "Bristol, UK, New York"
  - 50% don't have valid location information
    "Hogwarts, In the traffic..blah, Sitting On A Taco"

Problem Definition

- Input: A list of time and location stamps of the target posts.
- Output: f_{st} Intensity of target phenomenon at location s (e.g., New York) and time t (e.g., 0-1am)

Time	Location
2012-09-26 17:35:23	New York US
2012-09-27 12:17:52	N/A
2012-09-27 08:28:12	(-98.24, 23.22)
...	

		Time (*t*)		
		0-1am	1-2am	2-3am
Location (*s*)	California	f(1,1)	f(1,2)	f(1,3)
	New York	f(2,1)	f(2,2)	f(2,3)
	Washington	f(3,1)	f(3,2)	f(3,3)

Why Simple Estimation is Bad

- $f_{st} = x_{st}$, the count of target posts in bin (s, t)
- Justification: MLE of the model $x \sim \text{Poisson}(f)$

$$P(x) = \frac{f^x e^{-f}}{x!}$$

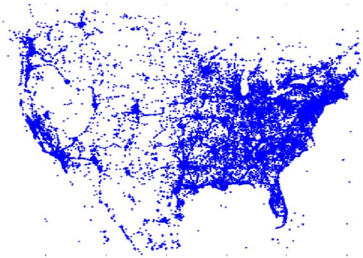
- However,
  - Population Bias: Assume $f_{st} = f_{s't'}$, if more users in (s, t), then $x_{st} > x_{s't'}$
  - Imprecise location: Posts without location stamp, noisy user profile location
  - Zero/Low counts: If we don't see tweets from Antarctica, no penguins there?

Correcting Population Bias

- Social media user activity intensity g_{st}

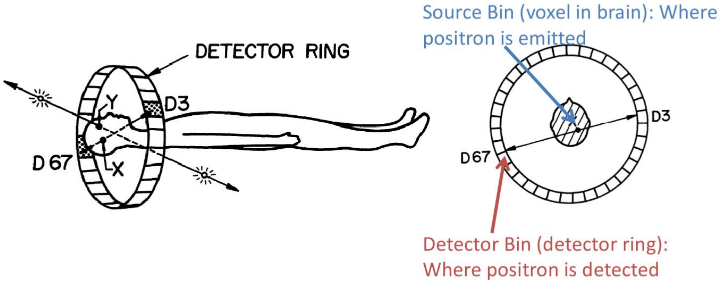
$$x \sim \mathrm{Poisson}(\eta(f, g))$$

- Link function (target post intensity) $\eta(f, g) = f \cdot g$
- Count of all posts $z \sim \mathrm{Poisson}(g)$
- g_{st} can be accurately recovered



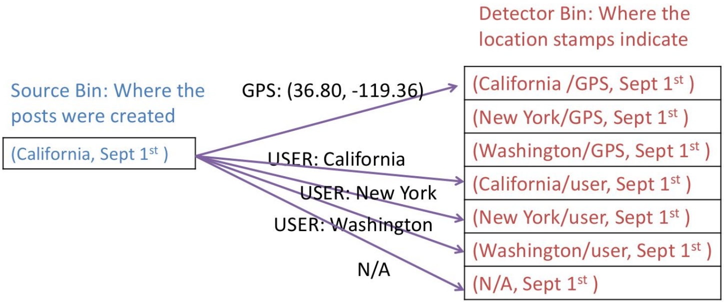
Handling Imprecise Location

Positron Emission Tomography (PET)



[Reproduced from Vardi et al(1985), A statistical model for positron emission tomography]

Handling Imprecise Location: Transition



Handling Imprecise Location: Transition

Fraction of posts with GPS coordinates

Probability that user was in California, but profile location is New York

.03	0	0
0	.03	0
0	0	.03
.37	.1	.01
.08	.3	.01
.02	.07	.45
.5	.5	.5

Fraction of posts without location stamps

X

Source Bin: Where the posts were created

- (California, Sept 1st)
- (New York, Sept 1st)
- (Washington, Sept 1st)

Intensity $\eta(f,g)$

=

Detector Bin: Where the location stamps indicate

- (California /GPS, Sept 1st)
- (New York/GPS, Sept 1st)
- (Washington/GPS, Sept 1st)
- (California/user, Sept 1st)
- (New York/user, Sept 1st)
- (Washington/user, Sept 1st)
- (N/A, Sept 1st)

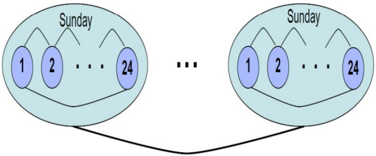
Intensity $h_i = \sum_{j=1}^{n} P_{ij}\eta(f_j, g_j)$

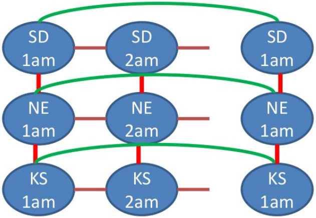
$$x_i \sim Poisson(h_i)$$

Handling Zero / Low Counts

Spatial Smoothness

Temporal Smoothness





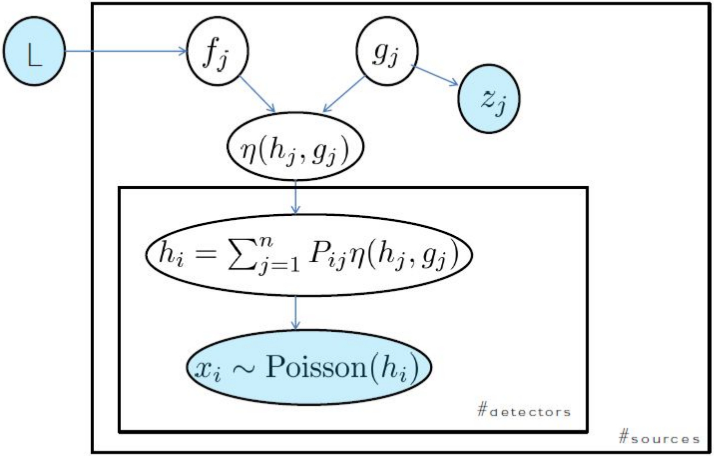
Weight Matrix W

$$D_{jj} = \sum_{k=1}^{n} W_{jk}$$

Graph Laplacian $L = D - W$

Regularizer $\Omega(f) = \frac{1}{2} \log f^T L \log f$

The Graphical Model



Optimization and Parameter Tuning

$$\min_{\theta \in \mathbb{R}^n} -\sum_{i=1}^{m} (x_i \log h_i - h_i) + \lambda \Omega(\theta)$$

$$\theta_j = \log f_j$$

$$h_i = \sum_{j=1}^{n} P_{ij} \eta(\theta_j, \psi_j)$$

- Quasi-Newton method (BFGS)
- Cross-Validation: Data-based and objective approach to regularization; Sub-sample events from the total observations

Theoretical Consideration

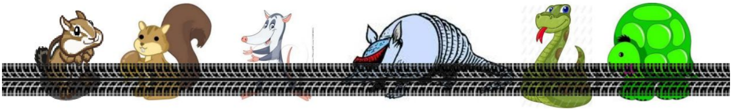
- How many posts do we need to obtain reliable recovery?
- If $x \sim \mathrm{Poisson}(h)$, then $\mathbb{E}[(\frac{x-h}{h})^2] = h^{-1} \approx x^{-1}$: more counts, less error
- Theorem: Let f be a Hölder α-smooth d-dimensional intensity function and suppose we observe N events from the distribution Poisson(f). Then there exists a constant $C_\alpha > 0$ such that

$$\inf_{\widehat{f}} \sup_{f} \frac{\mathbb{E}[\|\widehat{f} - f\|_1^2]}{\|f\|_1^2} \;\geq\; C_\alpha N^{\frac{-2\alpha}{2\alpha+d}}$$

- Best achievable recovery error is inversely proportional to N with exponent depending on the underlying smoothness

Roadkill Spatio-Temporal Intensity

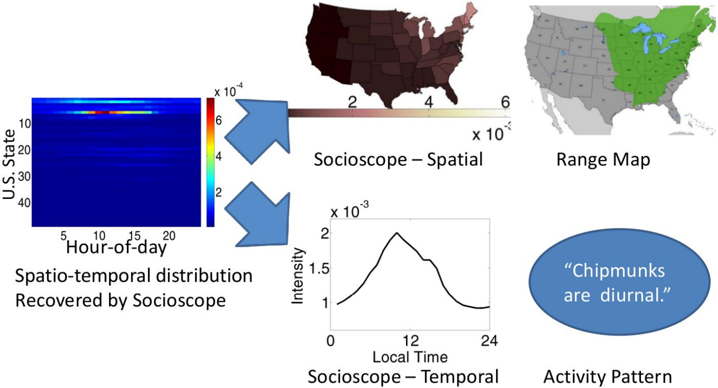
- The intensity of roadkill events within the continental US
- Spatio-Temporal resolution: State: 48 continental US states, hour-of-day: 24 hours
- Data source: Twitter
- Text classifier: Trained with 1450 labeled tweets. CV accuracy 90



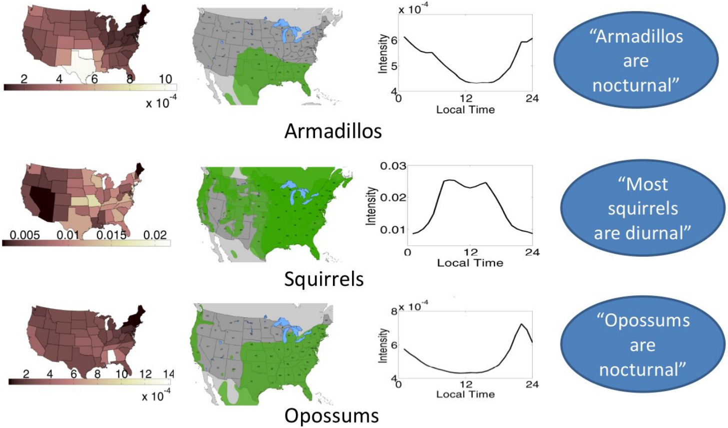
Text preprocessing

- Twitter streaming API: animal name + "ran over"
- Remove false positives by text classification
  "I almost ran over an armadillo on my longboard, luckily my cat-like reflexes saved me."
- Feature representation
  - Case folding, no stemming, keep stopwords
  - @john → @USERNAME, http://wisc.edu → HTTPLINK, keep #hashtags, keep emoticons
  - Unigrams + bigrams
- Linear SVM
  - Trained on 1450 labeled tweets outside study period
  - Cross validation accuracy 90%

Chipmunk Roadkill Results



Roadkill Results on Other Species



Outline

1. Spatio-Temporal Signal Recovery (Poisson Generative Model)

2. Finding Chatty Users (Multi-Armed Bandit)

Finding Chatty Users

- Find top k social media users on a topic
  - For example, via the bullying classifier [Xu,Jun,Zhu,Bellmore NAACL 2012]
- Trivial if we can monitor all users all the time
- But API only allows monitoring a small number (e.g. 5000) of users at a time
- Monitor each user "long enough?"

How Long is Long Enough for a Single User?

- Define a time slot (e.g., 1 hour)
- Define Boolean event
  - 1= the user posted anything on-topic in the time slot
  - 0= no post
- Define $p = Pr(\text{event=1})$
- Observe k time slots $X_1, \ldots, X_k \in \{0, 1\}$
- $\hat{p} = \frac{\sum_{i=1}^{k} X_i}{k}$
- How reliable is $\hat{p}$?

Hoeffding's Inequality

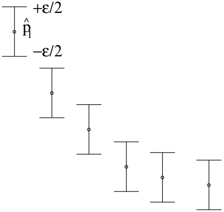
Let $X_1, \ldots, X_k$ be independent with $P(X_i \in [a,b]) = 1$ and the same mean p. Then for all $\epsilon > 0$,

$$P\left(\left| \frac{1}{k} \sum_{i=1}^{k} X_i - p \right| > \epsilon \right) \leq 2e^{-\frac{2k\epsilon^2}{(b-a)^2}}.$$

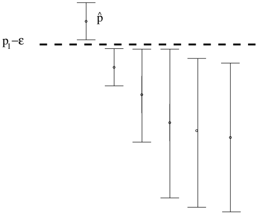
- We have $P\left(|\hat{p} - p| > \epsilon\right) \leq 2e^{-2k\epsilon^2}$.
- Define $\delta = 2e^{-2k\epsilon^2}$, then $\epsilon = \sqrt{\frac{\log \frac{2}{\delta}}{2k}}$ or $k = \frac{\log \frac{2}{\delta}}{2\epsilon^2}$.
- For any $\delta > 0$, with probability at least $1 - \delta$, $|\hat{p} - p| \leq \sqrt{\frac{\log \frac{2}{\delta}}{2k}}$.
- With $\frac{\log \frac{2}{\delta}}{2\epsilon^2}$ samples, with probability at least $1 - \delta$, $|\hat{p} - p| \leq \epsilon$.
- Confidence interval or Probably-Approximately-Correct (PAC) analysis

Uniform Monitoring is Wasteful

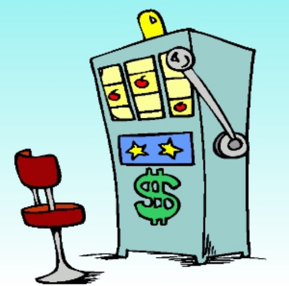
To find an ϵ-best arm $(p > p_1 - \epsilon)$ out of n arms, uniform monitoring needs a total of $O\left(\frac{n}{\epsilon^2} \log \frac{n}{\delta}\right)$ samples [Even-Dar et al. 2006]



Median Elimination (a Multi-Armed Bandit algorithm) needs $O\left(\frac{n}{\epsilon^2} \log \frac{1}{\delta}\right)$ samples



One-Armed Bandit



Expected reward $p \in [0, 1]$

Stochastic Multi-Armed Bandit Problem

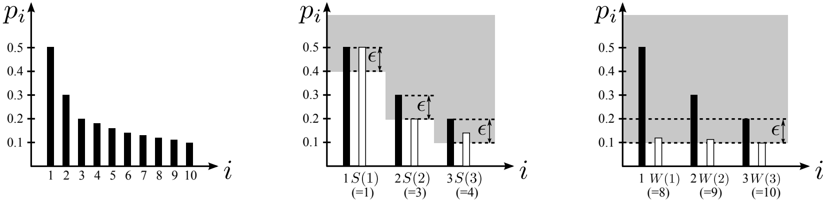
- Known parameters: number of arms n.
- Unknown parameters: n expected rewards $p_1 \geq \ldots \geq p_n \in [0, 1]$.
- For each round $t = 1, 2, \ldots$
  1. the learner chooses an arm $a_t \in \{1, \ldots, n\}$ to pull
  2. the world draws the reward $X_t \sim \mathrm{Bernoulli}(p_{a_t})$ independent of history
- The learner does not see the reward of non-chosen arms in each round.
- Pure exploration problem: find arms with the largest p's as quickly as possible.

Chatty Users as Multi-Armed Bandit

- User = arm, n = number of users (e.g. millions)
- p_i = user i's probability to post
- Monitoring user for a time slot = pulling that arm
- Reward X_t = did the user post anything?
- Pure exploration: with the least monitoring, find top-m users who post the most ($m \ll n$)
  - exactly the top-m users $p_1 \geq \ldots \geq p_m$, or
  - approximately the top-m users?

Approximate Top-m Arms: Strong vs. Weak Guarantee

- Let $S \subseteq \{1, \ldots, n\}$ and let $S(j)$ denote the arm whose expected reward is j-th largest in S.
- S is strong (ϵ, m)-optimal if $p_{S(i)} \geq p_i - \epsilon, \forall i = 1, ..., m$.
- S is weak (ϵ, m)-optimal if $p_{S(i)} \geq p_m - \epsilon, \forall i = 1, ..., m$.



arms — strong $(\epsilon, 3)$-optimal — weak $(\epsilon, 3)$-optimal

Multi-Armed Bandit Algorithms for Finding Top-m Arms

Guarantee	Algorithm	Sample Complexity (Worst Case)
Exact	SAR	$O(\frac{n}{\epsilon^2}(\log n)\left(\log\frac{n}{\delta}\right))$
Strong (ϵ, m)-optimal	EH	$O(\frac{n}{\epsilon^2}\log\left(\frac{m}{\delta}\right))$
Weak (ϵ, m)-optimal	Halving	$O(\frac{n}{\epsilon^2}\log\left(\frac{m}{\delta}\right))$
Weak (ϵ, m)-optimal	LUCB	$O(\frac{n}{\epsilon^2}\log\left(\frac{n}{\delta}\right))$

The Enhanced Halving (EH) Algorithm

Input: $n, m, \epsilon > 0, \delta > 0$

Output: m arms satisfying strong (ϵ, m)-optimality

$l \leftarrow 1, S_1 \leftarrow \{1, ..., n\}$, $n_1 \leftarrow n$, $\epsilon_1 \leftarrow \epsilon/4$, $\delta_1 \leftarrow \delta/2$

while $n_l > m$ **do**

$$n_{l+1} \leftarrow \begin{cases} \lceil n_l/2 \rceil & \text{if } |S_l| > 5m \\ m & \text{otherwise} \end{cases}$$

Pull every arm in S_l $\left\lceil \frac{1}{(\epsilon_l/2)^2} \log\left(\frac{5m}{\delta_l}\right) \right\rceil$ times

Compute $\widehat{p}_a^{(l)}, a \in S_l$, the empirical means from the sample drawn at iteration l

$S_{l+1} \leftarrow \{n_{l+1}$ arms with largest empirical means from $S_l\}$

$\epsilon_{l+1} \leftarrow \frac{3}{4}\epsilon_l, \delta_{l+1} \leftarrow \frac{1}{2}\delta_l$, $l \leftarrow l + 1$

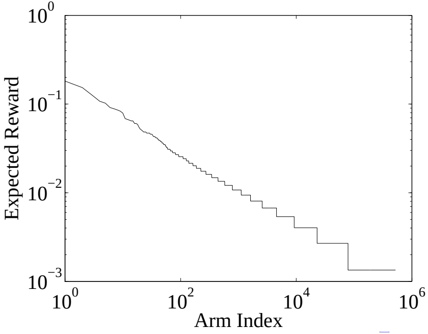
end while

Output $S := S_l$

Further improvement in constant: the Quantiling algorithm.

Application to Twitter Bullying

- $n = 522,062$ users, top $m = 100$
- 1 month total monitoring time (January 2013)
- $T = 31 \times 24 = 744$ time slots (pulls)
- Batch pulling 5000 arms at a time (*user streaming* API)
- Reward $X_{it} = 1$ if user i posts bullying-related tweets (judged by a text classifier) in time slot t.
- $log - log$ plot of expected reward follows the power law:



Experiments

Strong error of a set of m arms S:

$$\max_{i=1\ldots m} \{p_i - p_{S(i)}\}$$

the smallest ϵ with which S is strong (ϵ, m)-optimal.

Methods	Strong Error
EH	0.1040 ($\pm$ 0.004)
Quantiling	0.0478 ($\pm$ 0.002)
Halving	0.0999 ($\pm$ 0.004)
LUCB	0.1474 ($\pm$ 0.004)
LUCB/Batch	0.0826 ($\pm$ 0.004)
SAR	0.0678 ($\pm$ 0.003)
Uniform	0.0870 ($\pm$ 0.003)

Summary

- We present two social media mining tasks:
  - estimating intensity from counts
  - identifying the most chatty users
- Naive heuristic methods do not take full advantage of the data
- Mathematical models with provable properties extract knowledge from social media better.
- Acknowledgments

  - Collaborators: Amy Bellmore, Aniruddha Bhargava, Kwang-Sung Jun, Robert Nowak, Jun-Ming Xu
  - National Science Foundation IIS-1216758 and IIS-1148012, Global Health Institute at the University of Wisconsin-Madison.