Socioscope: Spatio-Temporal Signal Recovery from Social Media

Jun-Ming Xu†, Aniruddha Bhargava*, Robert Nowak*, Xiaojin Zhu†*
Department of †Computer Sciences, *Electrical and Computer Engineering, University of Wisconsin-Madison

When, Where, How Much

Many real-world phenomena can be represented by a spatio-temporal signal. Examples include wildlife mortality, algal booms, and hail damage. Direct instrumental sensing is often difficult and expensive.

Humans as Sensors

Social media offers a unique sensing opportunity for such signals, where users serve the role of "sensors" by posting their experiences of a target phenomenon. For instances, social media users readily post their encounters with dead animal: "*I saw a dead crow on its back in the middle of the road.*" (called target post)

Socioscope: Problem Definition

Input: A list of time and location stamps of the target posts.

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)

Output: Intensity of target phenomenon at the discrete spatiotemporal bins $f_{s,t}$

		Time		
		2012-09-26	2012-09-27	2012-09-28
Location	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)

Simple Estimation is Bad

$f_{s,t} = x_{s,t}$, the count of target posts in bin (s, t)

Justification: MLE of the model $x \sim Poisson(f)$

Population Bias :Even $f_{s,t} = f_{s',t'}$, if more users in (s, t) , then $x_{s,t} > x_{s',t'}$

Imprecise location: 3% have GPS coordinates, 47% have valid user profile location, 50% don't have GPS or profile location.

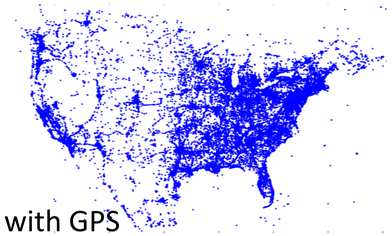
Zero/Low counts: If no tweeters in Antarctica, does it mean no penguins there?

Correcting Human Population Bias

The target post intensity depends on both target phenomenon intensity ***f*** and social media user activity intensity ***g***

$$x \sim Poisson(\eta(f,g))$$

Example: $\eta(f,g) = f.g$



g can be accurately recovered with all posts with GPS

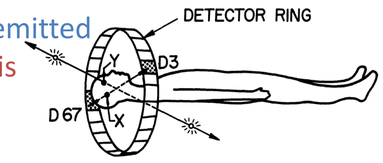
Handling Imprecise Location Stamps

Positron Emission Tomography

Source Bin (Voxel in Brain): Where positron is emitted

Detector Bin (Detector Ring): Where positron is detected

Transition probability from source bin to detector bin is determined by physical laws.



[Reproduced from Vardi *et al*(1985)

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

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

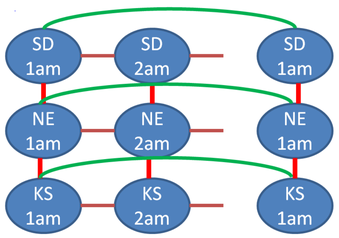
(California /GPS, Sept 1st)
(New York/GPS, Sept 1st)
(Washington/GPS, Sept 1st)
(California/user, Sept 1st)
(New York/user, Sept 1st)
(Washington/user, Sept 1st)
(N/A, Sept 1st)

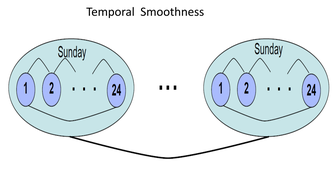
Intensity $h_i = \sum_{j=1}^{n} P_{ij}\eta(f_j, g_j)$

$x_i \sim Poisson(h_i)$

Handling Zero/Low Counts

Define graph-based regularizer to incorporate domain knowledge, such as smoothness, daily/weekly patterns.





Weight Matrix W Graph Laplacian $L = D - W$

$D_{jj} = \sum_{k=1}^{n} W_{jk}$ Regularizer $\Omega(f) = \frac{1}{2}\log f^T L \log f$

Case Study: Roadkill

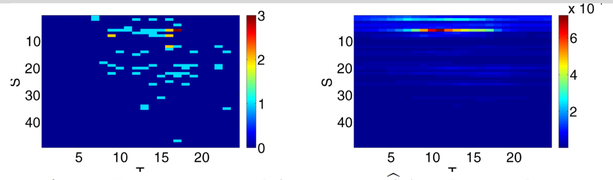
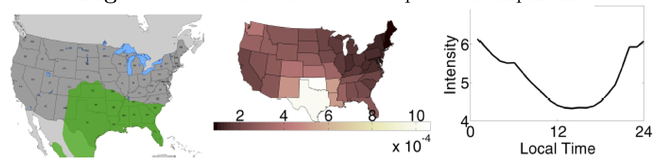
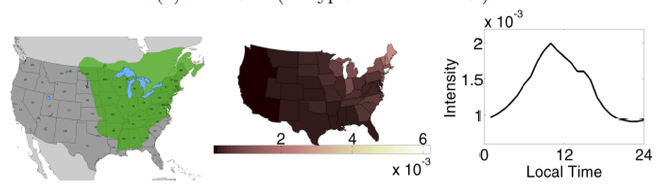


Fig. 4. Raw counts and Socioscope $\widehat{\mathbf{f}}$ for chipmunks



(a) armadillo (Dasypus novemcinctus)



(b) chipmunk (Tamias striatus)

Fig. 3. Socioscope estimates match animal habits well. (Left) range map from NatureServe, (Middle) Socioscope $\widehat{\mathbf{f}}$ aggregated spatially, (Right) $\widehat{\mathbf{f}}$ aggregated temporally.

Future Work

1. Include text classification confidence in input.

Text Classifier Confidence	Time	Location
0.9	2012-09-26 17:35:23	Wisconsin US
0.2	2012-09-26 17:38:33	N/A
0.6	2012-09-27 12:17:52	N/A

2. Handle the temporal delay and spatial displacement between target event and the generation of a posts "*So the pigeon I ran over yesterday must have some bird friends in high places. Car is full of bird shit*."
3. Incorporate psychology factors to better model the human sensors." For instance, a person may not bother to tweet about a chipmunk roadkill, but may be eager to do so upon seeing a moose roadkill.

We thank Megan K. Hines for her guidance on wildlife. This work is supported in part by Global Health Institute at the University of Wisconsin-Madison.

xujm@cs.wisc.edu