Social media data and computational models of mobility: A review for demography

Momin M. Malik[1] & Jürgen Pfeffer[1,2]
[1]Institute for Software Research, Carnegie Mellon University
[2]Bavarian School of Public Policy, Technical University of Munich

Tenth Annual AAAI Conference on Web and Social Media
Workshop on Social Media and Demographic Research
Cologne, Germany
17 May 2016

Slides available at:
mominmalik.com/smdr2016.pdf

Motivation

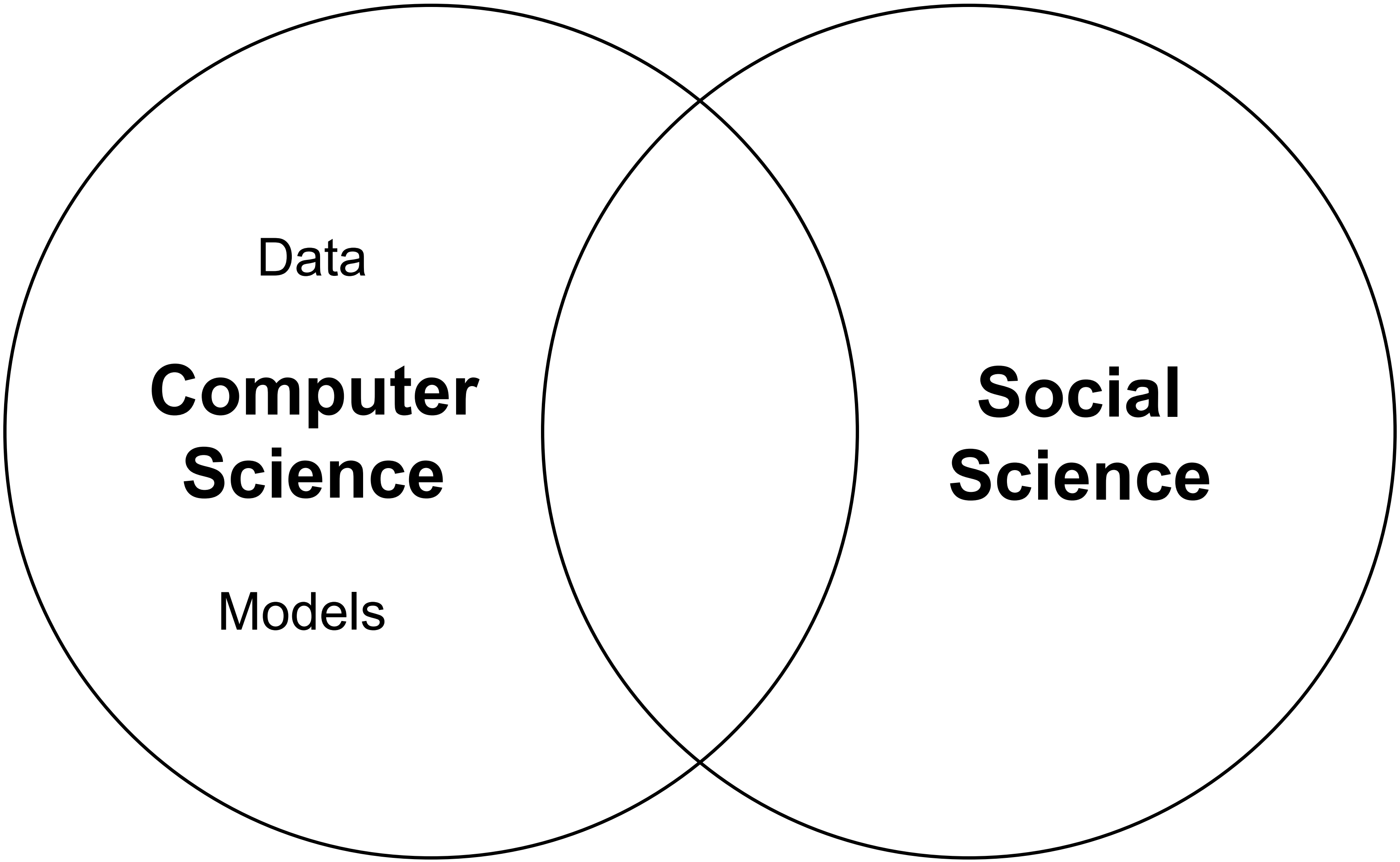
Concern:

- Industry is years ahead of social science in having access to data and computational expertise
- Industry has used opportunity to make enormous findings and advances (Savage & Burrows, 2007)

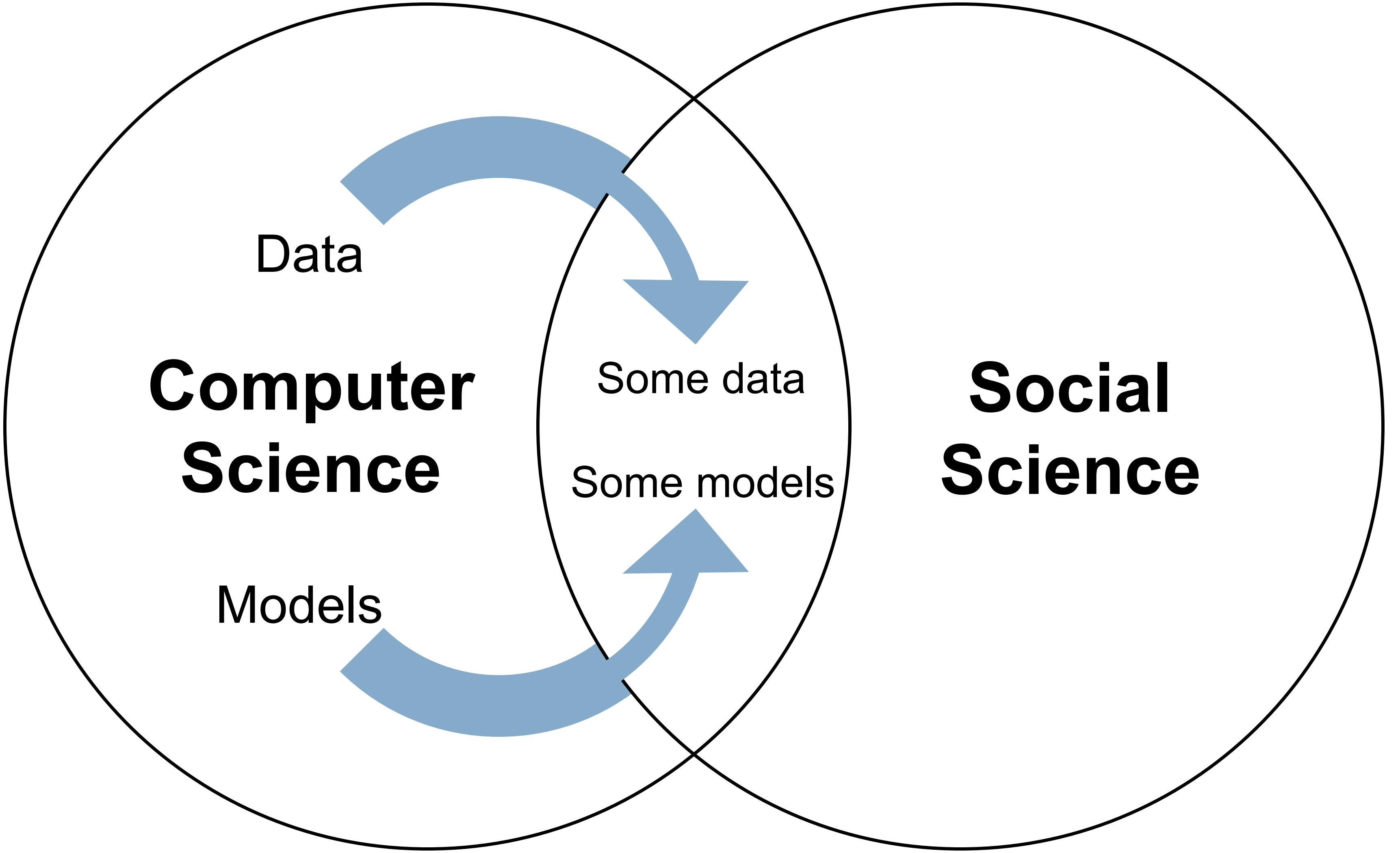
Reality:

- Models that 'work' in a commercial context may be quite uninteresting for academics (Burrows & Gane, 2006)
- Computational modeling (i.e., machine learning) focuses on *prediction*, not explanation (Shmueli, 2010; Breiman, 2001)*
- **Best-fitting model may not be "true"** (Shmueli, 2010)

Goal



Goal


Data
Computer Science
Models
Some data
Some models
Social Science

Tasks in CS/industry

- Inferring location from noisy data (map apps on phones)
- Recommender systems
  - For movies, people, shopping, restaurants, social events
- Geographic topic analysis
  - Words associated with regions "in order to enrich the functional description of locations for designing advanced location-based services" (Gao & Liu, 2015)
- Event Detection
  - Automate detection of natural disasters, sports events
- Simulation for testing
  - "Realistic" behavior to test technical infrastructure
- Location prediction
  - Forecasting demand, or recommendation systems again

Types of data

- GPS location logs
- Cell phone tower access logs
  - May be combined with call logs
- **Social media data:**
  - Anything that allows “check-ins”: Foursquare, subsets of Twitter, Facebook
  - Also known as:
    - Location-based social networks (LBSN)
    - Volunteered Geographic Information (VGI)

Social Media Data

Existing research into biases and problems

Problems

Problems are massive (Tufekci, 2014; Ruths & Pfeffer, 2014).

We must think about the context in which the data are generated! (van Dijck, 2013; Gehl, 2014)

- People sell bots to inflate metrics (Donath 2008); lots of spam (Thomas et al., 2011; 2013) makes data messy
- Idiosyncratic behaviors and conventions (boyd et al., 2010; Java et al., 2007; Kwak et al., 2010)
- Unreliable data access (Morstatter et al., 2013)
- Platform effects (see my talk tomorrow!)
- International differences (Poblete et al., 2011)
- Changes over time (Liu et al., 2014; van Dijck, 2013)

Representativeness

Work comparing Census data to Twitter data:

- Mislove et al. (2011): Uneven distribution in US based on self-identified location on Twitter
- Sloan et al. (2013): Gender distribution similar to UK Census
- Hecht and Stephens (2014): Bias in US geotagged tweets use towards urban areas
- Longle et al. (2015): Overrepresentation of young males, White British users in London geotagged tweets
- Malik et al. (2015): Overrepresentation of geotagged Tweets in block groups with young users, high Asian populations, black populations, Latino populations

Note: this work assumes that Census data are the "ground truth"!

Consequences

All these biases matter!

- Twitter opinion does not match public opinion (Mitchell & Hitlin, 2013): i.e., conclusions based on social media data are "wrong"
- Can correct for population (Zagheni & Weber 2015), but others?
- Even if models are fitted to social media/real-world correspondences, such correspondences (and hence the models) can break down under a slight change in context (Cohen & Ruths, 2013, "Classifying Political Orientation on Twitter: It's not Easy!")
- **"Prediction" is a technical term that means "fitted values:" a model that "predicts" well is actually just a model that *fits* well.** Model fit is a heuristic for future performance, not a guarantee (Gayo-Avello 2013; 2012a; 2012b); and a lack of causal understanding makes good future performance less likely

Alternative: Social media data as a "test bed"

Video, "Tracking Malte Spitz":
https://www.youtube.com/watch?v=J1EKvWot-3c

Malte Spitz / Die Zeit / Future Journalism Project Media Lab, 2010

Places to apply?

$$
\begin{aligned}
Population_{t+1} = Population_t \\
&+ (Births_t - Deaths_t) \\
&+ (Immigration_t - Emigration_t)
\end{aligned}
$$

Places to apply?

$$\begin{aligned} \mathit{Population}_{t+1} = \mathit{Population}_t \\ + (\mathit{Births}_t - \mathit{Deaths}_t) \\ + (\mathit{Immigration}_t - \mathit{Emigration}_t) \end{aligned}$$

How can we better characterize migration? Are there already relevant models in computer science?

(Thanks to Ridhi Kashyap, Katharina Kinder-Kurlanda)

Mobility Models

What does computer science/engineering do around mobility? How does it all work?

Characterizing mobility

- Currently, we found no models that take continuous paths and use them to create “mobility profiles”
- Most models, at some point, discretize or make bins (image: Bayir et al., 2009)

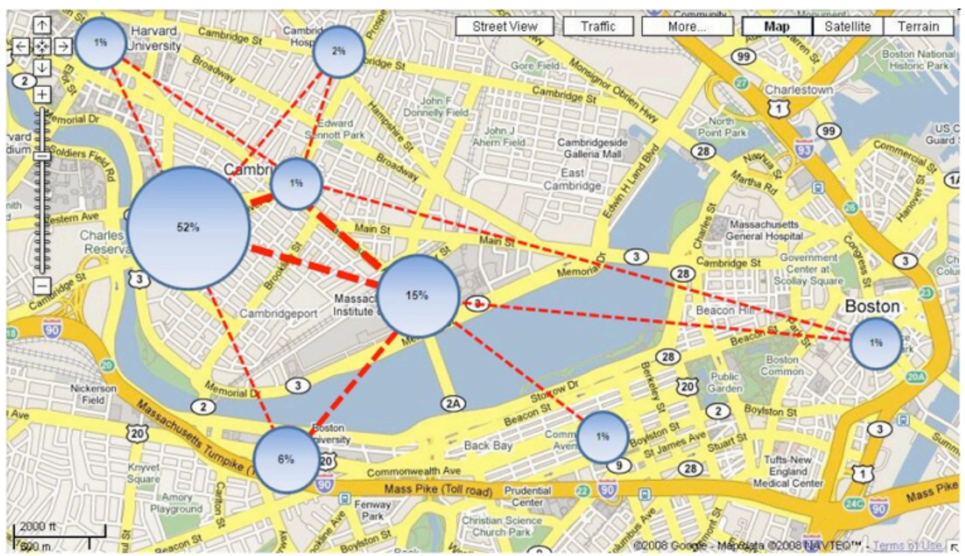
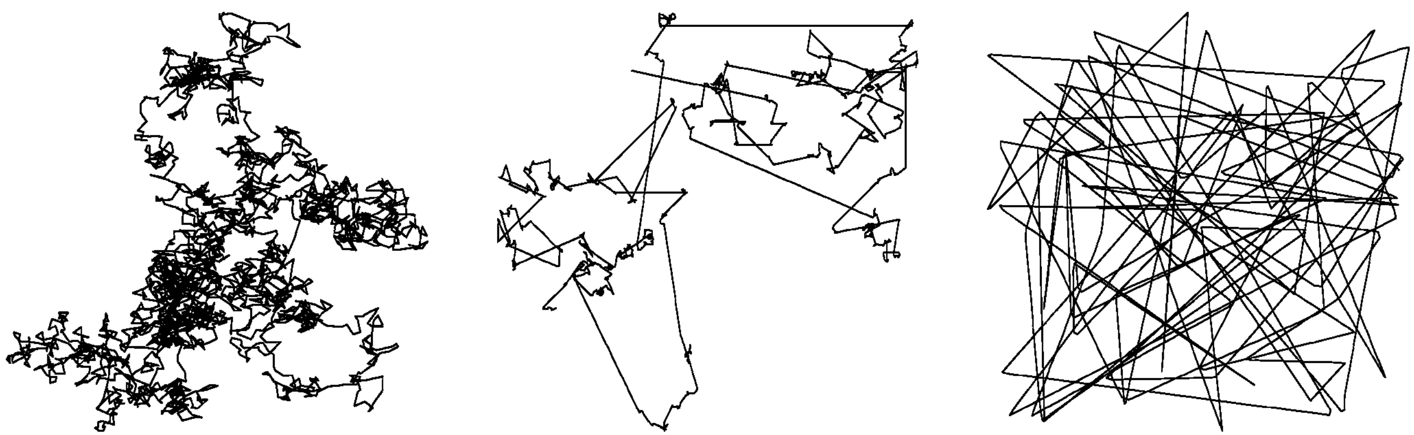


Figure 10: Time distribution for end locations on map for user X

Basic task: generating "realistic" behavior

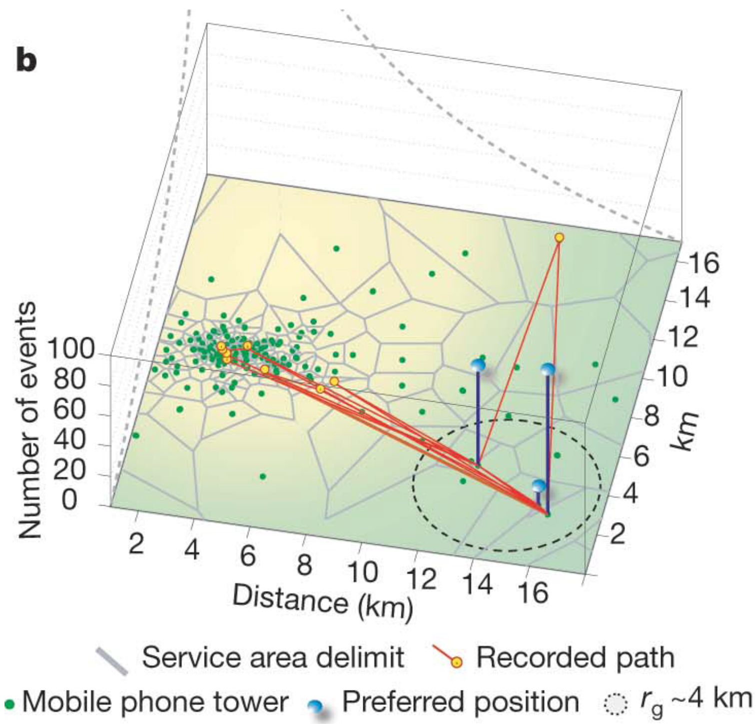
- Simplest model is that of a "random walk" process
  - Unrealistic
- Can also use models for describing particles, "Brownian motion;" still not realistic
- "Lévy walks" are between the two (image: Rhee et al., 2011)



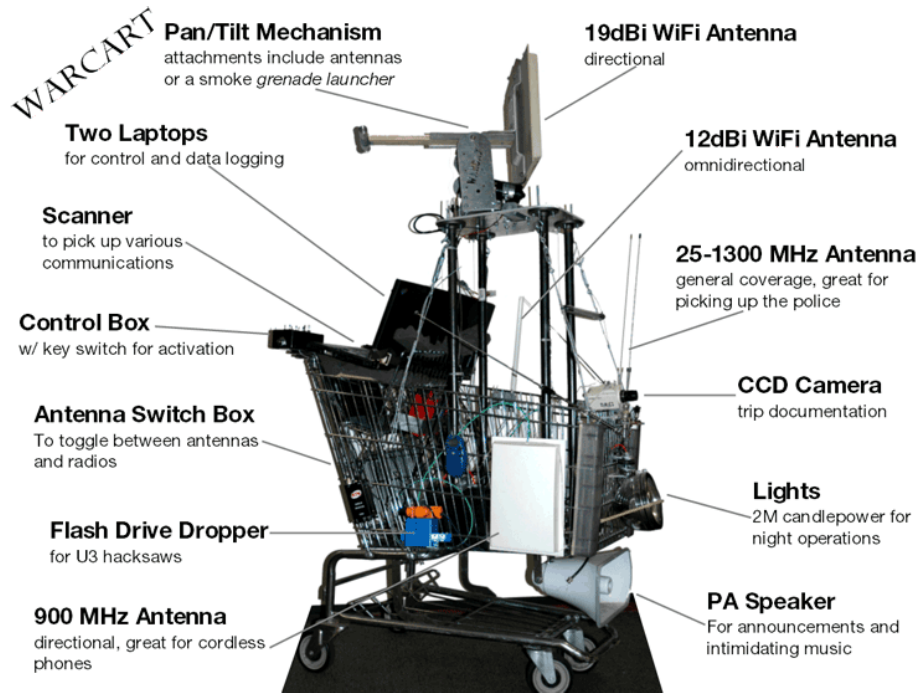
(a) (b) (c)

Fig. 1. Sample trajectories of (a) BM, (b) Levy walk, and (c) RWP.

Inferring location



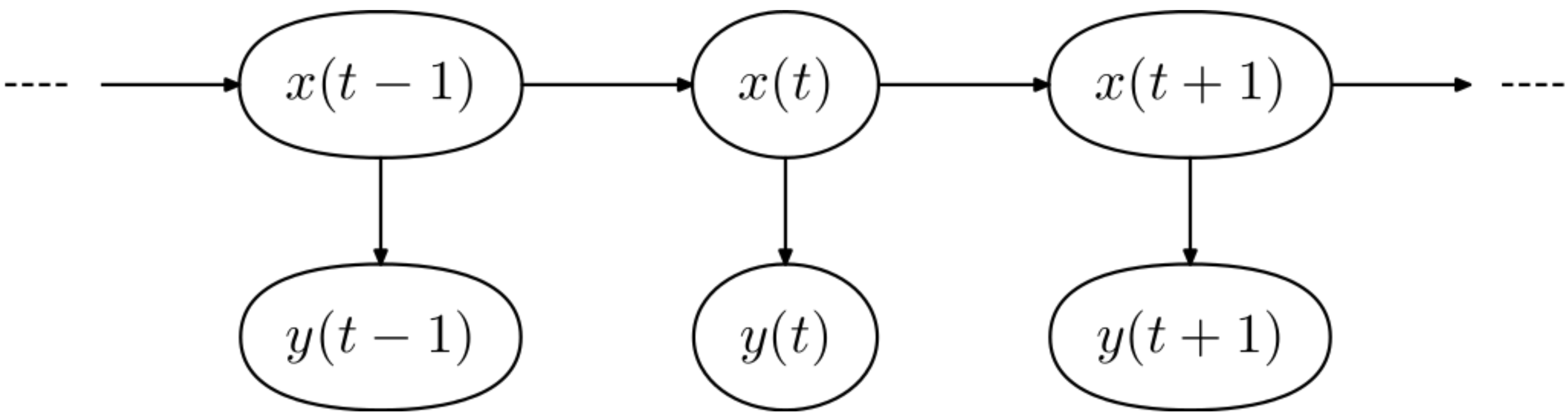
González et al., 2008



Zac Anderson, “Warcart”

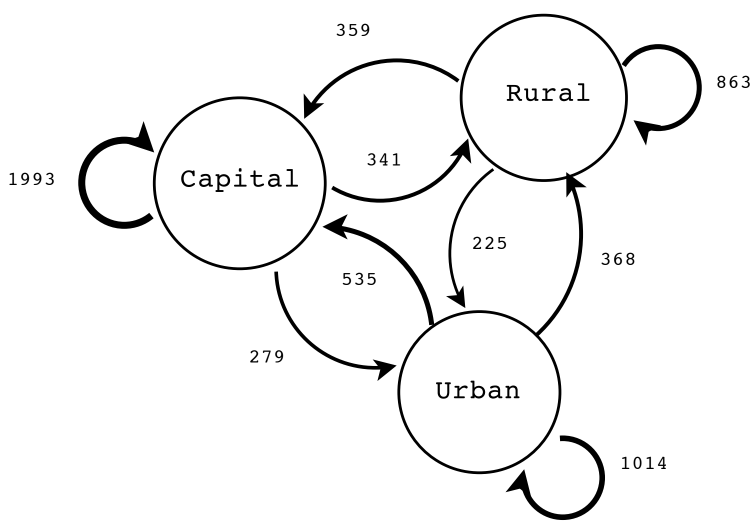
Inferring *trajectories*

- Can infer trajectories from noisy point data
- Use a *Markov Model* that represents transitions between states
- (Keywords: Hidden Markov Model, State Space Model, Kalman Filter)



Transition Matrix

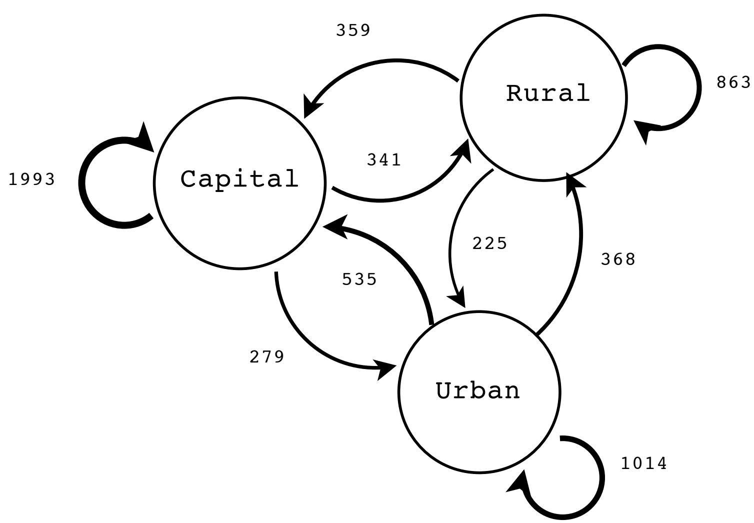
- Key component of Markov Models is the *transition matrix*, which represents transitions between states
- States can be locations as well! (image: Eagle et al., 2009)



from \ to	Capital	Rural	Urban
Capital	1993	341	279
Rural	359	863	225
Urban	535	368	1014

Transition Matrix

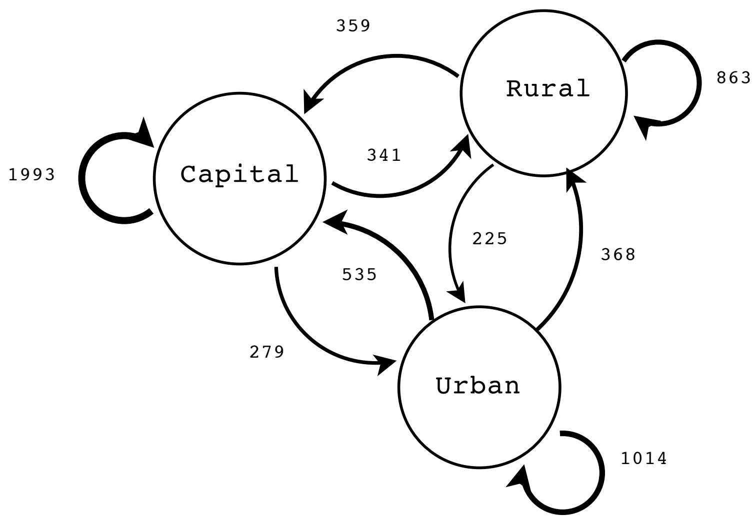
- Interpret entry *ij* as from row *i* to column *j*.
- Row-normalize counts (divide each row by the row sum)
- Normalization gives frequencies, an estimate of probabilities



to / from	Capital	Rural	Urban
Capital	1993/2613	341/2613	279/2613
Rural	359/1447	863/1447	225/1447
Urban	535/1917	368/1917	1014/1917

Transition Matrix

- Interpret entry *ij* as <u>from</u> row *i* <u>to</u> column *j*.
- Row-normalize counts (divide each row by the row sum)
- Normalization gives frequencies, an estimate of probabilities



from \ to	Capital	Rural	Urban
Capital	.763	.131	.107
Rural	.248	.596	.155
Urban	.279	.192	.529

Conclusions

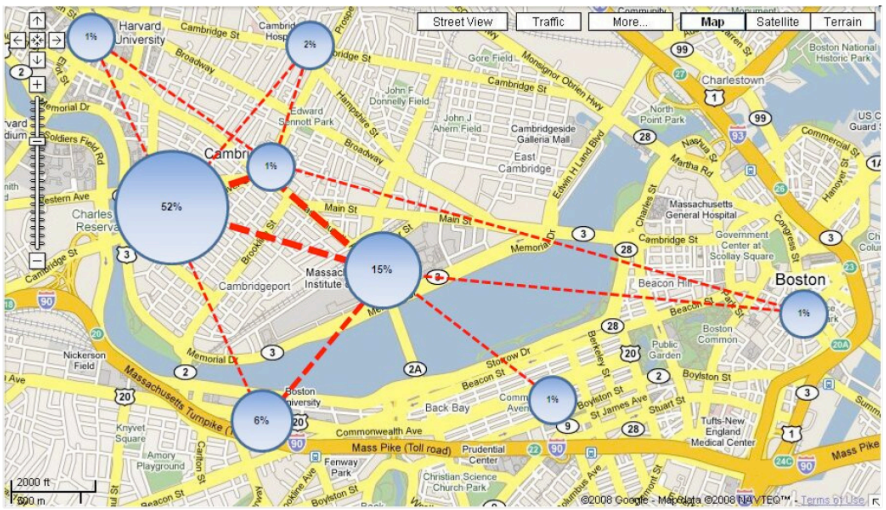


Figure 10: Time distribution for end locations on map for user X

- Be careful before using social media data!
- Good news: Social science is not replaced
- Bad news: Social science may have little to contribute to the goals of CS
- There are representations in computer science (and statistics) that may be very useful for demography: specifically, transition matrices

Thank you! Questions?

momin.malik@cs.cmu.edu
http://mominmalik.com/smdr2016.pdf

References

Bayir, Demirbas, Eagle 2009 Discovering spatiotemporal mobility profiles of cellphone users https://dspace.mit.edu/handle/1721.1/53746

boyd, Golder, Lotan 2010 Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5428313

Breiman 2001 Statistical modeling: The two cultures https://projecteuclid.org/euclid.ss/1009213726

Burrows & Gane 2006 Geodemographics, software and class http://soc.sagepub.com/content/40/5/793.abstract

Cohen & Ruths 2013 Classifying political orientation on Twitter: It's not easy! http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6128

Donath 2008 Signals in social supernets http://onlinelibrary.wiley.com/doi/10.1111/j.1083-6101.2007.00394.x/full

Eagle, de Montjoye, Bettencourt 2009 Community computing: Comparisons between rural and urban societies using mobile phone data http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5284288

Gao & Liu 2015 Mining human mobility in location-based social networks http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=7110013

Gayo-Avello 2012a 'I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper': A balanced survey on election prediction using Twitter data http://arxiv.org/abs/1204.6441

References

Gayo-Avello 2012b No, you cannot predict elections with Twitter http://ieeexplore.ieee.org/xpl/abstractAuthors.jsp?arnumber=6355554

Gayo-Avello 2012c Don't turn social media into another 'literary digest' poll http://cacm.acm.org/magazines/2011/10/131406-dont-turn-social-media-into-another-literary-digest-poll/abstract

Gayo-Avello 2013 A meta-analysis of state-of-the-art electorial prediction from Twitter data http://ssc.sagepub.com/content/31/6/649.short

Gehl 2014 Reverse engineering social media: Software, culture, and political economy in new media capitalism http://www.temple.edu/tempress/titles/2275_reg.html

González, Hidalgo, Barabási 2008 Understanding individual human mobility patterns http://www.nature.com/nature/journal/v453/n7196/full/nature06958.html

Hecht & Stephens 2014 A tale of cities: Urban biases in volunteered geographic information http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8114

Java, Song, Finin, Tseng 2007 Why we Twitter: Understanding microblogging usage and communities http://dl.acm.org/citation.cfm?id=1348556

Kwak et al 2010 What is Twitter, a social network or a news media? http://an.kaist.ac.kr/traces/WWW2010.html

References

Liu, Kliman-Silver, Mislove 2014 The tweets they are a-changin': Evolution of Twitter users and behavior http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8043/0

Longley, Adnan, Lansley 2015 The geotemporal demographics of Twitter usage http://epn.sagepub.com/content/47/2/465.abstract

Malik, Lamba, Nakos, Pfeffer 2015 Population bias in geotagged tweets http://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10662

Mitchell & Hitlin, 2013 Twitter reaction to events often at odds with overall public opinion http://www.pewresearch.org/2013/03/04/twitter-reaction-to-events-often-at-odds-with-overall-public-opinion/

Morstatter, Pfeffer, Liu, Carley 2013 Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6071

Problete, Garcia, Mendoza, Jaimes 2011 Do all birds tweet the same? Characterizing Twitter around the world http://dl.acm.org/citation.cfm?id=2063724

Rhee et al 2011 On the Levy-talk nature of human mobility http://netsrv.csc.ncsu.edu/export/infocom2008_mobility_final.pdf

Ruths & Pfeffer 2014 Social media for large studies of behavior http://science.sciencemag.org/content/346/6213/1063

References

Savage & Burrows 2007 The coming crisis of empirical sociology, http://soc.sagepub.com/content/41/5/885.abstract

Shmueli 2010 To explain or to predict? https://projecteuclid.org/euclid.ss/1294167961

Sloan et al 2013 Knowing the tweeters: Deriving sociologically relevant demographics from Twitter http://www.socresonline.org.uk/18/3/7.html

Thomas et al 2011 Suspended accounts in retrospect: An analysis of Twitter spam http://conferences.sigcomm.org/imc/2011/docs/p243.pdf

Thomas, McCoy, Paxson 2013 Trafficking fraudulent accounts: The role of the underground market in Twitter spam and abuse https://www.usenix.org/conference/usenixsecurity13/technical-sessions/paper/thomas

Tufekci 2014 Big questions for social media big data: Representativeness, validity, and other methodological pitfalls http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8062

Van Dijck 2013 The culture of connectivity: A critical history of social media http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199970773.001.0001/acprof-9780199970773

Zagheni & Weber 2015 Demographic research with non-representative internet data http://www.emeraldinsight.com/doi/abs/10.1108/IJM-12-2014-0261