From Twitter API to Social Science Paper

Presentation for the ICOS Big Data Boot Camp

Todd Schifeling

5/22/14

Outline

I. Collecting Twitter Data with a Snowball

II. Motivation for Collecting the Data
   i. Big Data-Social Science Divide
   ii. Possible Solutions

Snowballing Twitter Data

Procedure:

- starting point
- network search
- selection principle

Snowballing Twitter Data

Procedure:

- starting point: **Scratchtruck**
- network search: **friends**
- selection principle: **self-description matches 2 dictionaries**

Twitter Data Calls

- friends.ids returns friendship ties (from, to)
  - 5000 per call at one minute per call = 5000 friendship ties per minute (but only one user per minute)
- users.lookup returns user info (name, description, location, last tweet, etc.)
  - 100 per call at six seconds per call = 1000 users per minute

more info at https://dev.twitter.com/docs/api/1.1

Snowballing Twitter Data

Results:

Steps	Time	Possible	Already Done	Selected	Collected	Friends
1	1 min	1	0	1	1	3002
2	1 hr 42 mins	3002	0	91	88	106769
3	3 dys 4 hrs 24 mins	67764	2383	4359	4324	2511143

Workflow for Food Trucks Paper

- Get Twitter data on possible trucks
- Identify trucks
- Get idiosyncratic trucks from Twitter via in-degree
- Match trucks to cities
- Get additional data (demographics, chains, microbreweries, weather, etc.)
- Regressions!

Co-author: Daphne Demetry, Northwestern University

Now We're Doing Social Science!

Table 2 - Negative Binomial Regression Models Predicting the Number of Gourmet Food Trucks Created in a City, *N* = 287

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
National chains (%)		-8.84***							-6.464***	-6.501***
		(1.381)							(1.487)	(1.495)
Breweries			0.15***						0.106**	0.109**
			(0.039)						(0.041)	(0.035)
Farmer's markets				0.081***					0.005	
				(0.024)					(0.02)	
Creative workers (%)					5.963**				0.052	
					(2.091)				(2.295)	
College graduates (%)						0.03***			0.018*	0.018*
						(0.008)			(0.009)	(0.007)
Hi-income houses (%)							-0.008			
							(0.012)			
Racial HHI								-2.6***	-2.8***	-2.83***
								(0.713)	(0.695)	(0.684)
Total population	0.000**	0.000**	0.000*	0.000**	0.000**	0.000**	0.000**	0.000**	0.000*	0.000*
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Population density	0.000	0.000***	0.000	0.000	0.000	0.000	0.000	0.000	0.000**	0.000**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Rental costs	0.002***	0.002***	0.003***	0.003***	0.002***	0.002***	0.002***	0.002***	0.002***	0.002***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)
Extreme temp. rate	-0.71***	-0.502***	-0.689***	-0.659***	-0.72***	-0.696***	-0.705***	-0.722***	-0.518***	-0.519***
	(0.137)	(0.091)	(0.119)	(0.116)	(0.152)	(0.158)	(0.133)	(0.118)	(0.076)	(0.075)
Constant	-0.542	3.05***	-1.288**	-1.173*	-2.195**	-1.382**	-0.571	1.237	2.978**	3.038***
	(0.513)	(0.786)	(0.411)	(0.572)	(0.694)	(0.539)	(0.52)	(0.754)	(1.000)	(0.833)
Degrees of freedom	4	5	5	5	5	5	5	5	10	8
Wald χ^2	82.88	174.11	112.17	83.68	98.3	87.01	86.23	131.88	263.17	248.54

Note: Robust standard errors clustered around metropolitan areas are in parentheses.

*p≤.05; **p≤.01; ***p≤.001.

But Why Collect Twitter Data on Gourmet Food trucks?

How Well Do They Mesh?

	Social Science	Big Data
Measurement	fidelity	large unobtrusive N

How Well Do They Mesh?

	Social Science		Big Data
Measurement	fidelity	IDEAL	large unobtrusive N

How Well Do They Mesh?

	Social Science		Big Data
Measurement	fidelity	IDEAL	large unobtrusive N
Sampling	random		digital breadcrumbs

How Well Do They Mesh?

	Social Science		Big Data
Measurement	fidelity	IDEAL	large unobtrusive N
Sampling	random	CHASM	digital breadcrumbs

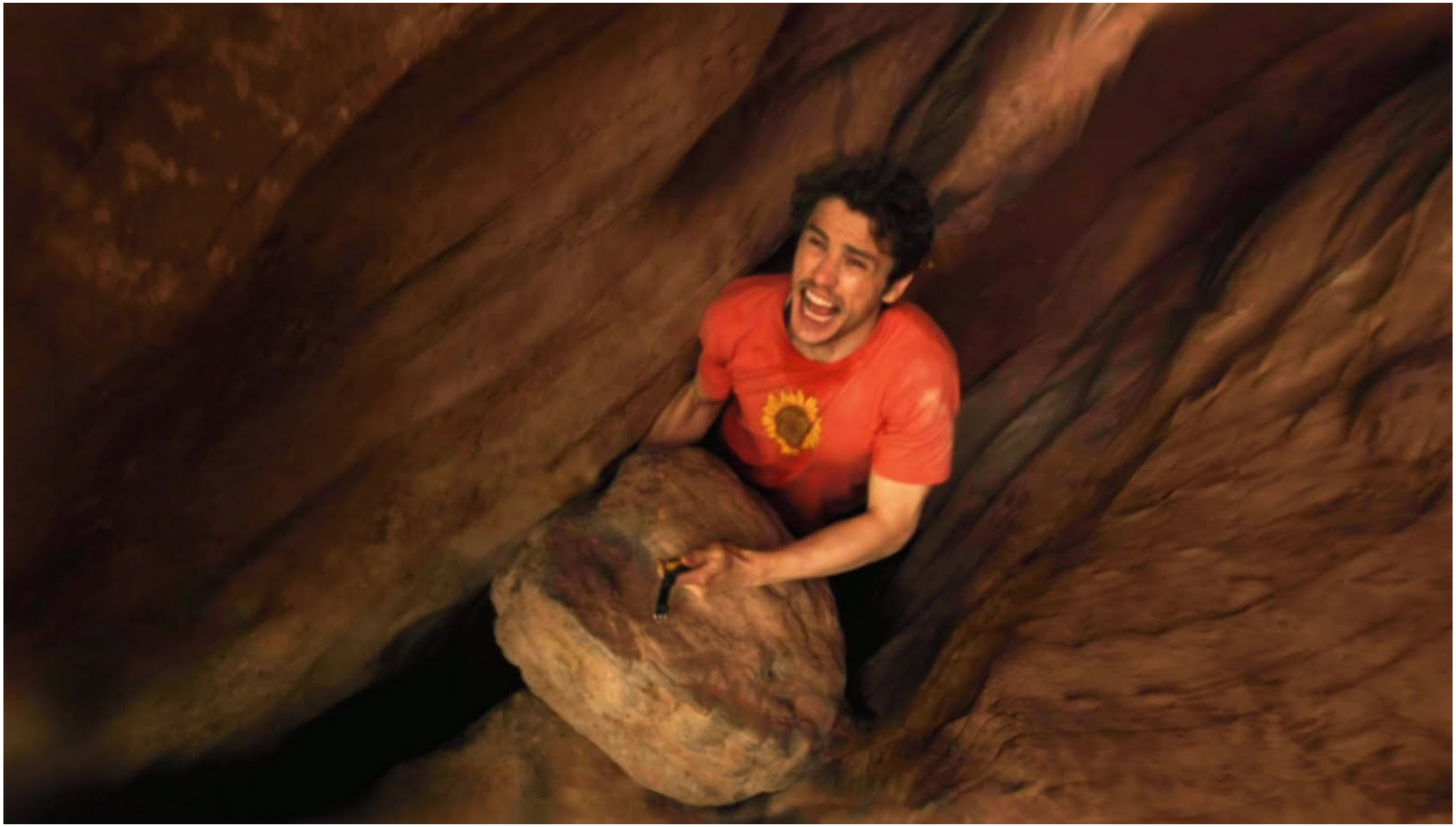
How Well Do They Mesh?

	Social Science		Big Data
Measurement	fidelity	IDEAL	large unobtrusive N
Sampling	random	CHASM	digital breadcrumbs
Causality	realism		description

How Well Do They Mesh?

	Social Science		Big Data
Measurement	fidelity	IDEAL	large unobtrusive N
Sampling	random	CHASM	digital breadcrumbs
Causality	realism	CHASM	description

The Fallout



A Possible Way Forward

Identify populations that simultaneously inhabit both offline and online worlds…

…which links sampling frames to available breadcrumbs, and ‘real’ to digital phenomena

A Typology of Examples that Cross the Offline/Online Divide

1. Offline activities that are more common online or are difficult to observe offline:
   - rare or deviant subcultures
   - bullying, deception, and other bad behaviors

A Typology of Examples that Cross the Offline/Online Divide

2. Offline activities with a significant online share:

- dating markets
- reviews of restaurants, books, movies, consumer goods, etc.
- neighborhood activism

A Typology of Examples that Cross the Offline/Online Divide

3. Offline activities that are also born online:
   - crowdsourcing projects
   - modern political ads
   - start-ups

Why the Case of Gourmet Food Trucks Bridges Offline and Online

- A new organizational form
- Twitter is crucial to the operations of the trucks
- Golden breadcrumbs get left behind

Comparison of Twitter Data to Standard Organizational Data

- Advantages: user-generated data, unfiltered by mediating data collector, digital breadcrumbs tracks organizational activity, relational data

- Disadvantages: less systematic comparison across organizations, have to clean and validate data yourself