Social Media as Social Science Data

Steven L. Wilson

V-Dem Institute, University of Gothenburg

From Tiananmen to Tahrir

Social Media as Social Science Data

- Theory
- My data collection project
- Sample projects
- Case study

Social Media as Social Science Data

Approach 1: social media as a causal variable in and of itself:

- Effect of social media on authoritarian regimes (Mazaid et al 2011, Tufeki and Wilson 2012, Farrell 2012, Tucker 2013)
- Effect of social media on mobilization/protest (Wilson 2016)

Social Media as Social Science Data

Approach 2: social media as a tool to measure something else:

- Detect effect of natural disasters (Sakashi, et al. 2010; Vieweg et al. 2010)
- Information diffusion through population (Lerman and Ghosh 2010, Romero et al 2011, Jansen et al 2009)
- Measure real-time effect of debates on voters (Diakopoulos and Shamma 2010)

Social Media as Social Science Data

Idealist claim: everything we want to know about public opinion is available via social media.

Limitation: that information is inaccessible because we don’t know what the denominator is.

The Problems

- The Representativeness Problem
  - Social media users not representative
  - *Vocal* social media users not representative
- The Population Problem
  - Might not even be *in* the population in question

The Problems

- The Representativeness Problem
  - Social media users not representative
  - *Vocal* social media users not representative
- The Population Problem
  - Might not even be *in* the population in question

– These are *not new* problems in social science

Social Media as Pseudo-Polls

- User meta data to correct sampling
- Semantic analysis to identify intervening variables (education, wealth, ethnicity)
- Survey sample of social media users to control for intervening variables

Data Sources

- Social Networks (Facebook)
- Blogs
- Microblogs (Twitter)

- Problems:
  - Country variation
  - Temporal variation
  - Public/private divide

Twitter Framework for Social Science

- Geocoded social media data can target populations
- Two approaches:
  - Context based
  - GPS based
- The additional problem: selecting on the dependent variable:
  - By time
  - By keyword

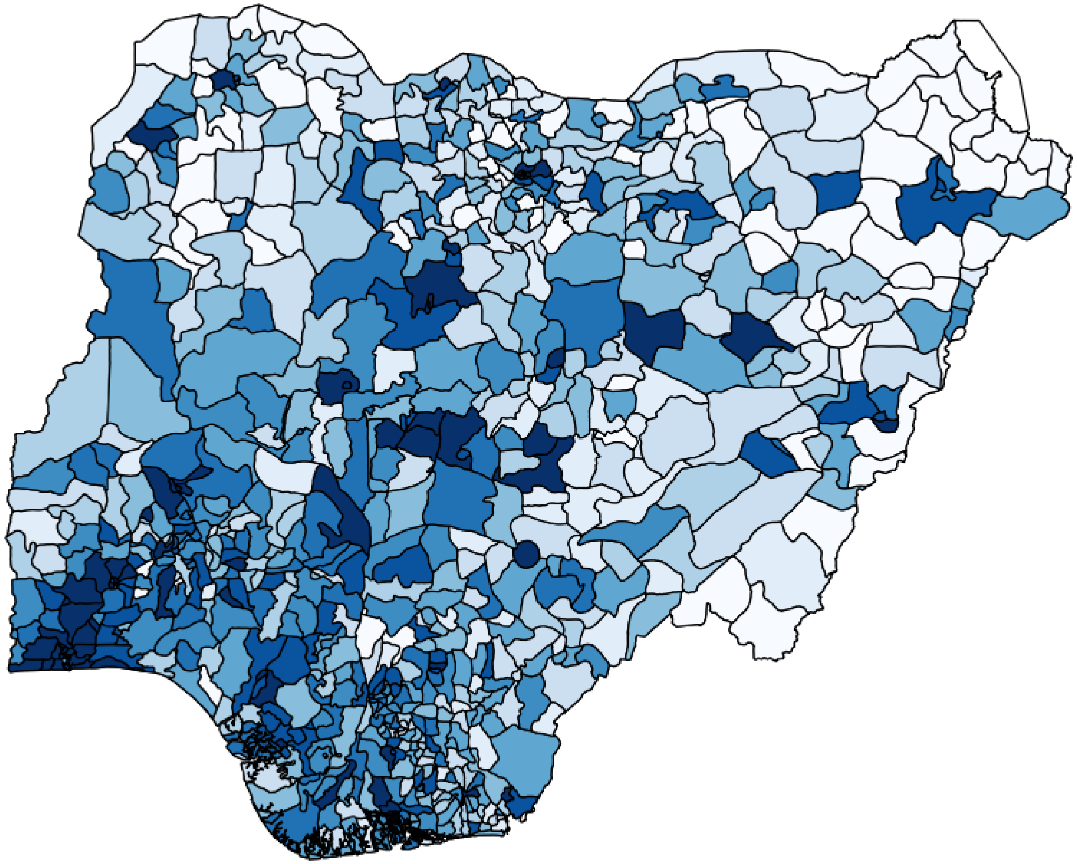
Twitter Framework for Social Science

- The Data:
  - Latitude and longitude
  - Full text and keywords
  - Metadata: followers, time zone, language

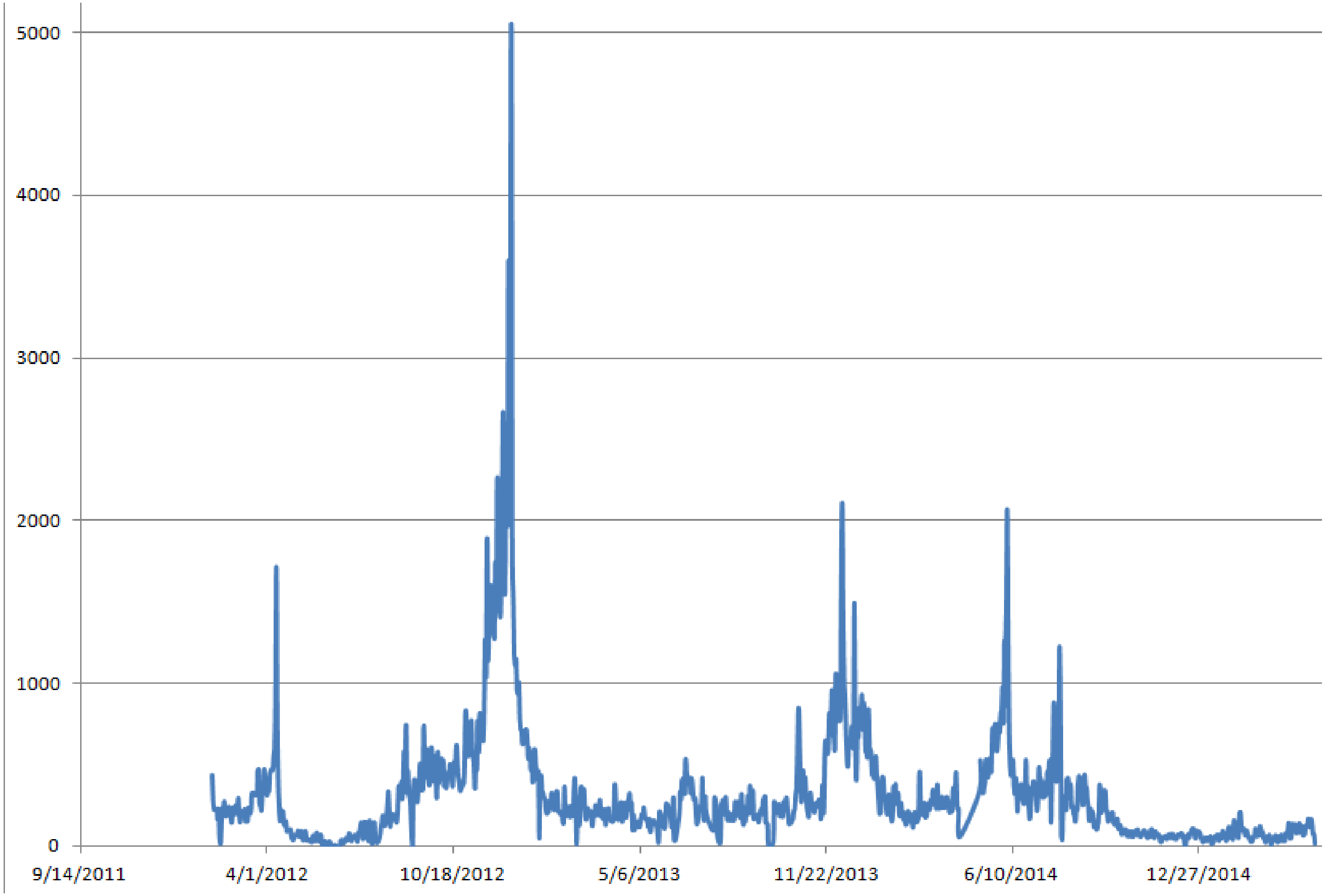
Twitter Framework for Social Science

- Every geocoded tweet (3 billion):
  - E. Europe, Asia, Africa (since Jan 2012)
  - Latin America (since Dec 2014)
- Keyword collection on demand:
  - Worldwide
  - Real time sub-national categorization

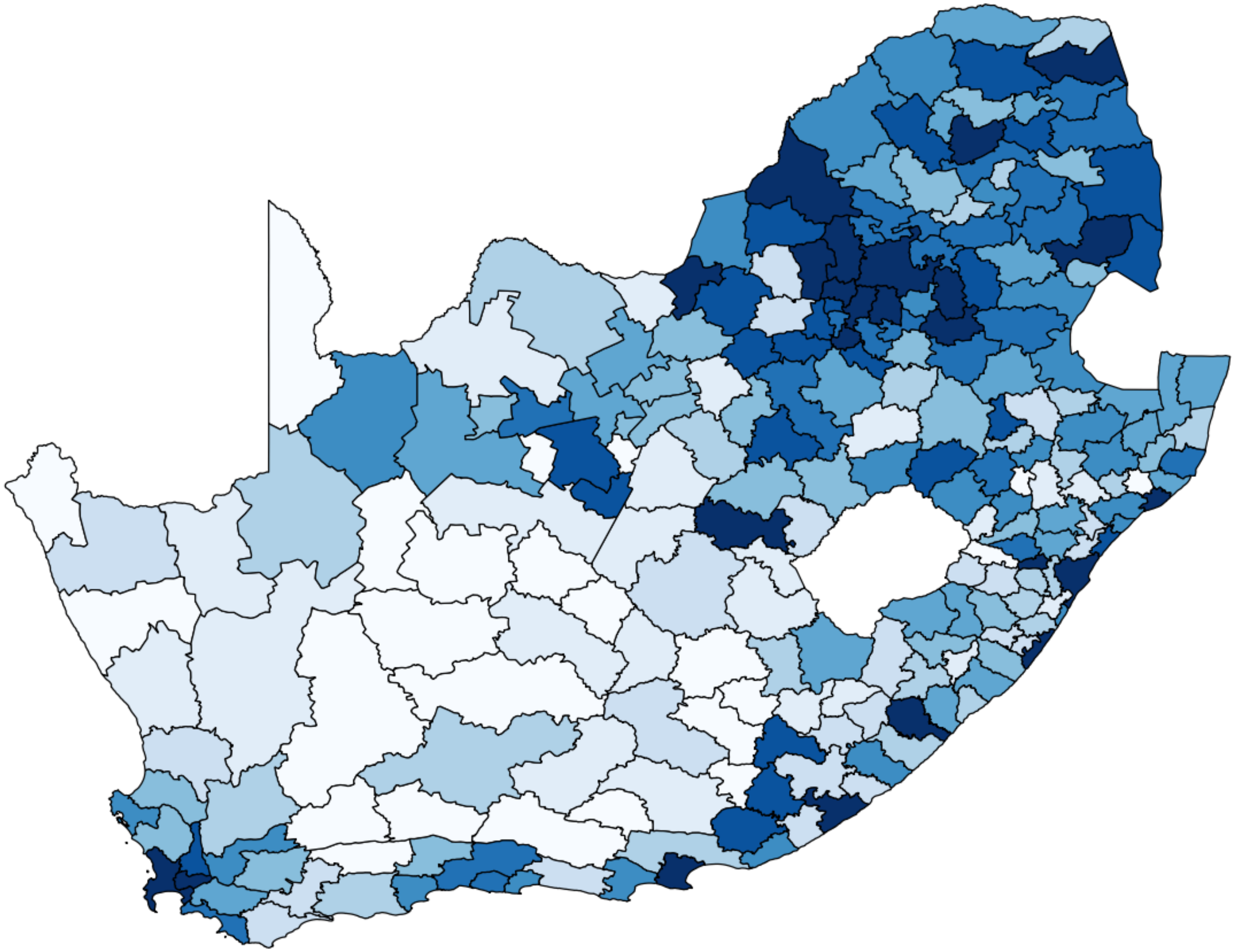
Project: Nigeria and Ethnicity



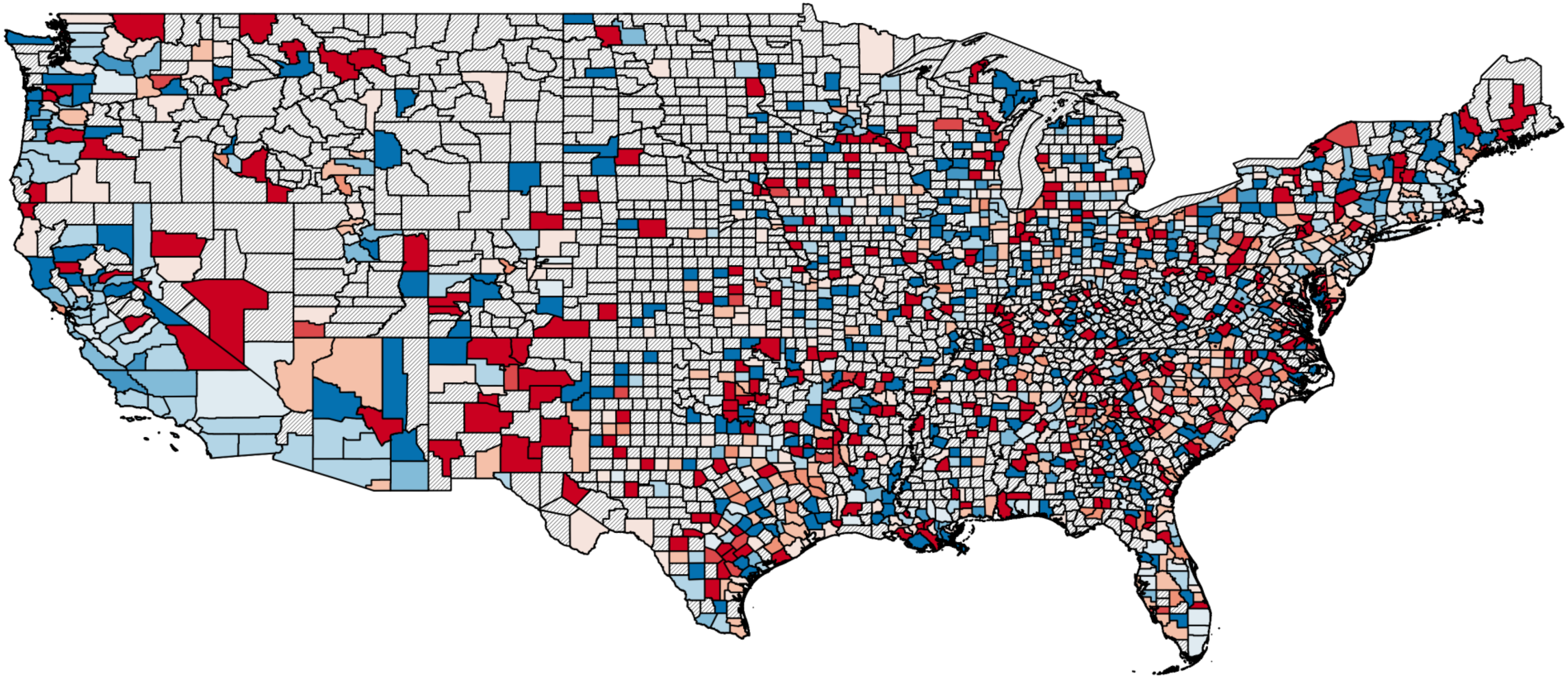
Project: South Korean Censorship



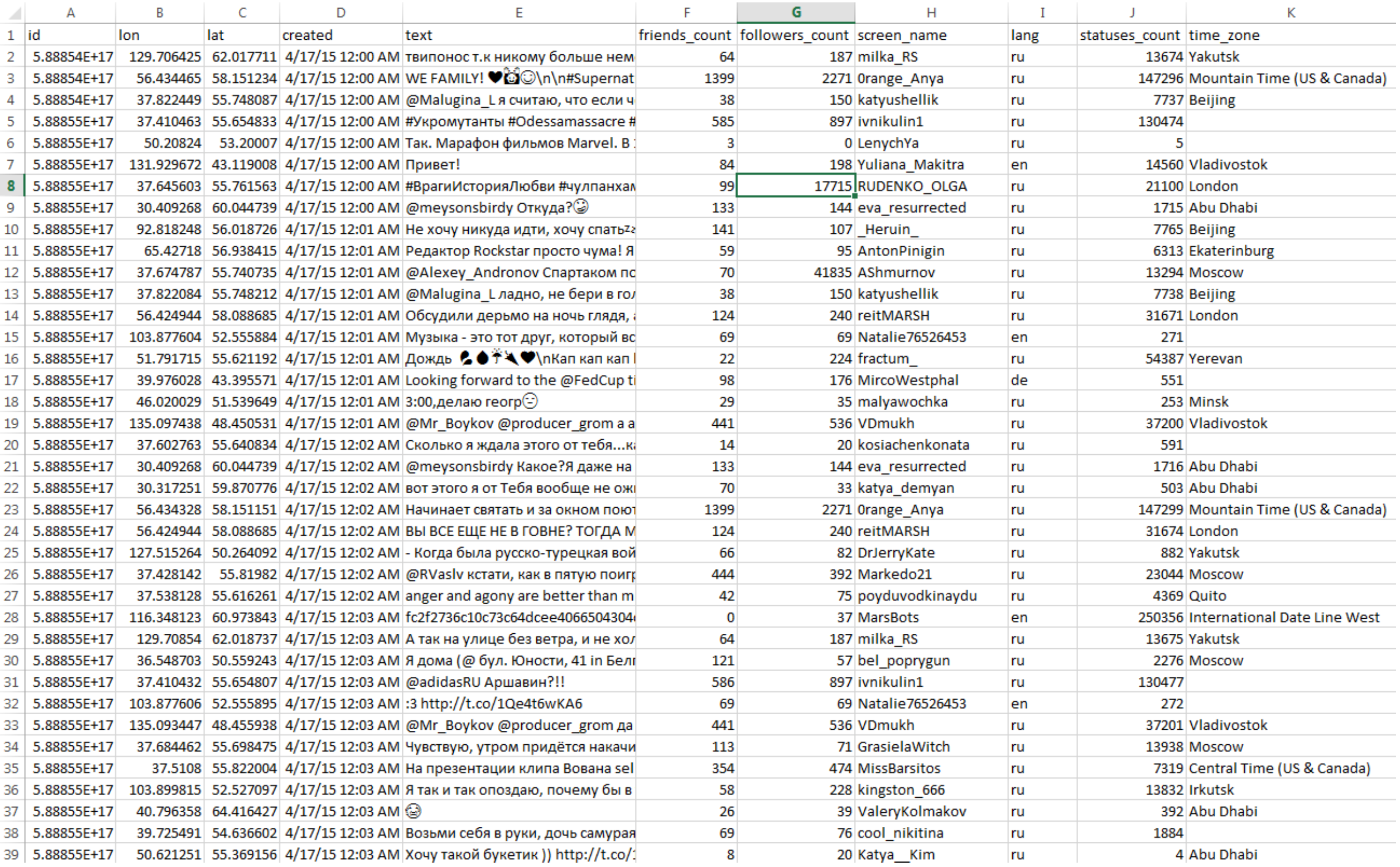
Project: South Africa #mustfall



Project: Latino vs. Hispanic

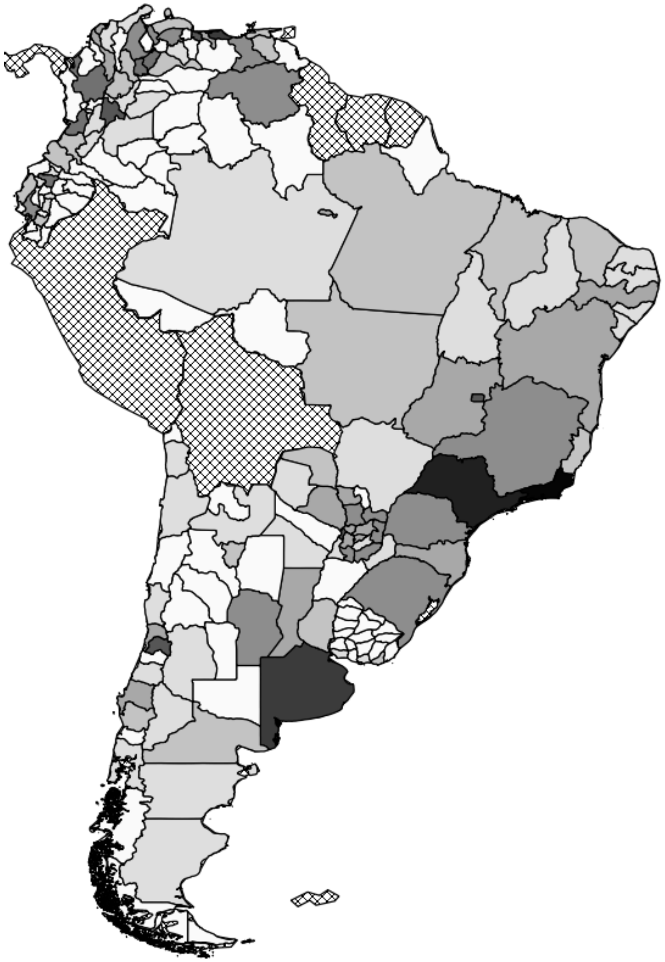


Project: Russia, “Direct Line”



	A	B	C	D	E	F	G	H	I	J	K
1	id	lon	lat	created	text	friends_count	followers_count	screen_name	lang	statuses_count	time_zone
2	5.88854E+17	129.706425	62.017711	4/17/15 12:00 AM	твипонос т.к никому больше нем	64	187	milka_RS	ru	13674	Yakutsk
3	5.88854E+17	56.434465	58.151234	4/17/15 12:00 AM	WE FAMILY! ♥👪☺\n\n#Supernat	1399	2271	0range_Anya	ru	147296	Mountain Time (US & Canada)
4	5.88854E+17	37.822449	55.748087	4/17/15 12:00 AM	@Malugina_L я считаю, что если ч	38	150	katyushellik	ru	7737	Beijing
5	5.88855E+17	37.410463	55.654833	4/17/15 12:00 AM	#Укромутанты #Odessamassacre #	585	897	ivnikulin1	ru	130474	
6	5.88855E+17	50.20824	53.20007	4/17/15 12:00 AM	Так. Марафон фильмов Marvel. В	3	0	LenychYa	ru	5	
7	5.88855E+17	131.929672	43.119008	4/17/15 12:00 AM	Привет!	84	198	Yuliana_Makitra	en	14560	Vladivostok
8	5.88855E+17	37.645603	55.761563	4/17/15 12:00 AM	#ВрагиИсторияЛюбви #чулпанхам	99	17715	RUDENKO_OLGA	ru	21100	London
9	5.88855E+17	30.409268	60.044739	4/17/15 12:00 AM	@meysonsbirdy Откуда?😜	133	144	eva_resurrected	ru	1715	Abu Dhabi
10	5.88855E+17	92.818248	56.018726	4/17/15 12:01 AM	Не хочу никуда идти, хочу спать💤	141	107	_Heruin_	ru	7765	Beijing
11	5.88855E+17	65.42718	56.938415	4/17/15 12:01 AM	Редактор Rockstar просто чума! Я	59	95	AntonPinigin	ru	6313	Ekaterinburg
12	5.88855E+17	37.674787	55.740735	4/17/15 12:01 AM	@Alexey_Andronov Спартаком пс	70	41835	AShmurnov	ru	13294	Moscow
13	5.88855E+17	37.822084	55.748212	4/17/15 12:01 AM	@Malugina_L ладно, не бери в го.	38	150	katyushellik	ru	7738	Beijing
14	5.88855E+17	56.424944	58.088685	4/17/15 12:01 AM	Обсудили дерьмо на ночь глядя,	124	240	reitMARSH	ru	31671	London
15	5.88855E+17	103.877604	52.555884	4/17/15 12:01 AM	Музыка - это тот друг, который вс	69	69	Natalie76526453	en	271	
16	5.88855E+17	51.791715	55.621192	4/17/15 12:01 AM	Дождь 💦💧☔🌂♥\nКап кап кап	22	224	fractum_	ru	54387	Yerevan
17	5.88855E+17	39.976028	43.395571	4/17/15 12:01 AM	Looking forward to the @FedCup ti	98	176	MircoWestphal	de	551	
18	5.88855E+17	46.020029	51.539649	4/17/15 12:01 AM	3:00,делаю геогр😒	29	35	malyawochka	ru	253	Minsk
19	5.88855E+17	135.097438	48.450531	4/17/15 12:01 AM	@Mr_Boykov @producer_grom a a	441	536	VDmukh	ru	37200	Vladivostok
20	5.88855E+17	37.602763	55.640834	4/17/15 12:02 AM	Сколько я ждала этого от тебя...к	14	20	kosiachenkonata	ru	591	
21	5.88855E+17	30.409268	60.044739	4/17/15 12:02 AM	@meysonsbirdy Какое?Я даже на	133	144	eva_resurrected	ru	1716	Abu Dhabi
22	5.88855E+17	30.317251	59.870776	4/17/15 12:02 AM	вот этого я от Тебя вообще не ож	70	33	katya_demyan	ru	503	Abu Dhabi
23	5.88855E+17	56.434328	58.151151	4/17/15 12:02 AM	Начинает святать и за окном поют	1399	2271	0range_Anya	ru	147299	Mountain Time (US & Canada)
24	5.88855E+17	56.424944	58.088685	4/17/15 12:02 AM	ВЫ ВСЕ ЕЩЕ НЕ В ГОВНЕ? ТОГДА М	124	240	reitMARSH	ru	31674	London
25	5.88855E+17	127.515264	50.264092	4/17/15 12:02 AM	- Когда была русско-турецкая вой	66	82	DrJerryKate	ru	882	Yakutsk
26	5.88855E+17	37.428142	55.81982	4/17/15 12:02 AM	@RVaslv кстати, как в пятую поигр	444	392	Markedo21	ru	23044	Moscow
27	5.88855E+17	37.538128	55.616261	4/17/15 12:02 AM	anger and agony are better than m	42	75	poyduvodkinaydu	ru	4369	Quito
28	5.88855E+17	116.348123	60.973843	4/17/15 12:03 AM	fc2f2736c10c73c64dcee406650430	0	37	MarsBots	en	250356	International Date Line West
29	5.88855E+17	129.70854	62.018737	4/17/15 12:03 AM	А так на улице без ветра, и не хо.	64	187	milka_RS	ru	13675	Yakutsk
30	5.88855E+17	36.548703	50.559243	4/17/15 12:03 AM	Я дома (@ бул. Юности, 41 in Бел	121	57	bel_poprygun	ru	2276	Moscow
31	5.88855E+17	37.410432	55.654807	4/17/15 12:03 AM	@adidasRU Аршавин?!!	586	897	ivnikulin1	ru	130477	
32	5.88855E+17	103.877606	52.555895	4/17/15 12:03 AM	:3 http://t.co/1Qe4t6wKA6	69	69	Natalie76526453	en	272	
33	5.88855E+17	135.093447	48.455938	4/17/15 12:03 AM	@Mr_Boykov @producer_grom да	441	536	VDmukh	ru	37201	Vladivostok
34	5.88855E+17	37.684462	55.698475	4/17/15 12:03 AM	Чувствую, утром придётся накачи	113	71	GrasielaWitch	ru	13938	Moscow
35	5.88855E+17	37.5108	55.822004	4/17/15 12:03 AM	На презентации клипа Вована sel	354	474	MissBarsitos	ru	7319	Central Time (US & Canada)
36	5.88855E+17	103.899815	52.527097	4/17/15 12:03 AM	Я так и так опоздаю, почему бы в	58	228	kingston_666	ru	13832	Irkutsk
37	5.88855E+17	40.796358	64.416427	4/17/15 12:03 AM	😓	26	39	ValeryKolmakov	ru	392	Abu Dhabi
38	5.88855E+17	39.725491	54.636602	4/17/15 12:03 AM	Возьми себя в руки, дочь самурая	69	76	cool_nikitina	ru	1884	
39	5.88855E+17	50.621251	55.369156	4/17/15 12:03 AM	Хочу такой букетик)) http://t.co/	8	20	Katya__Kim	ru	4	Abu Dhabi

Project: Latin American Protest



Project: Euromaidan

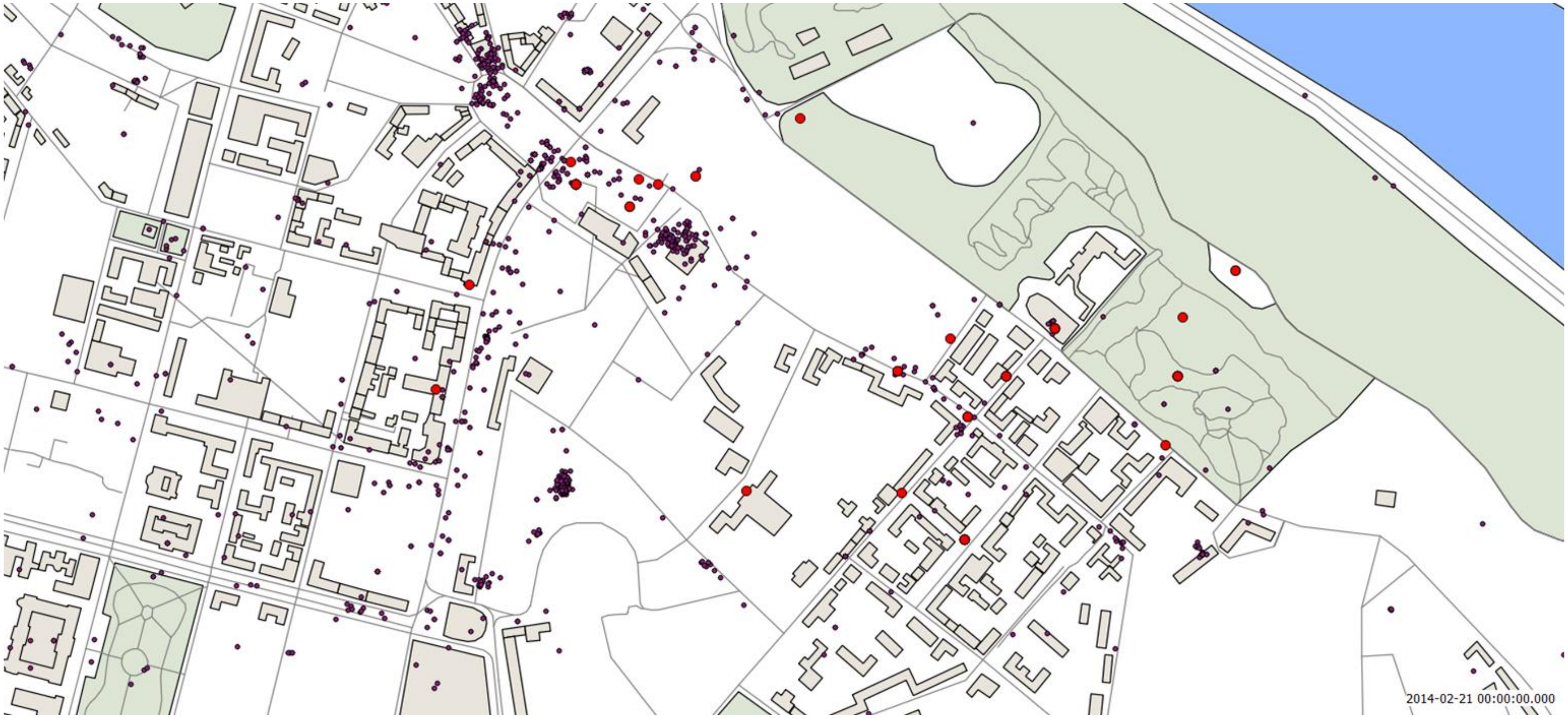
Operationalization: Tweets

- Tweets within Ukraine
  - 6 months of tweets (10/1/13 to 3/31/14)
  - 2.2 million tweets
- Operationalization:
  - Custom GIS code to identify raion
  - Number of friends (for weighting)

Operationalization: Protest

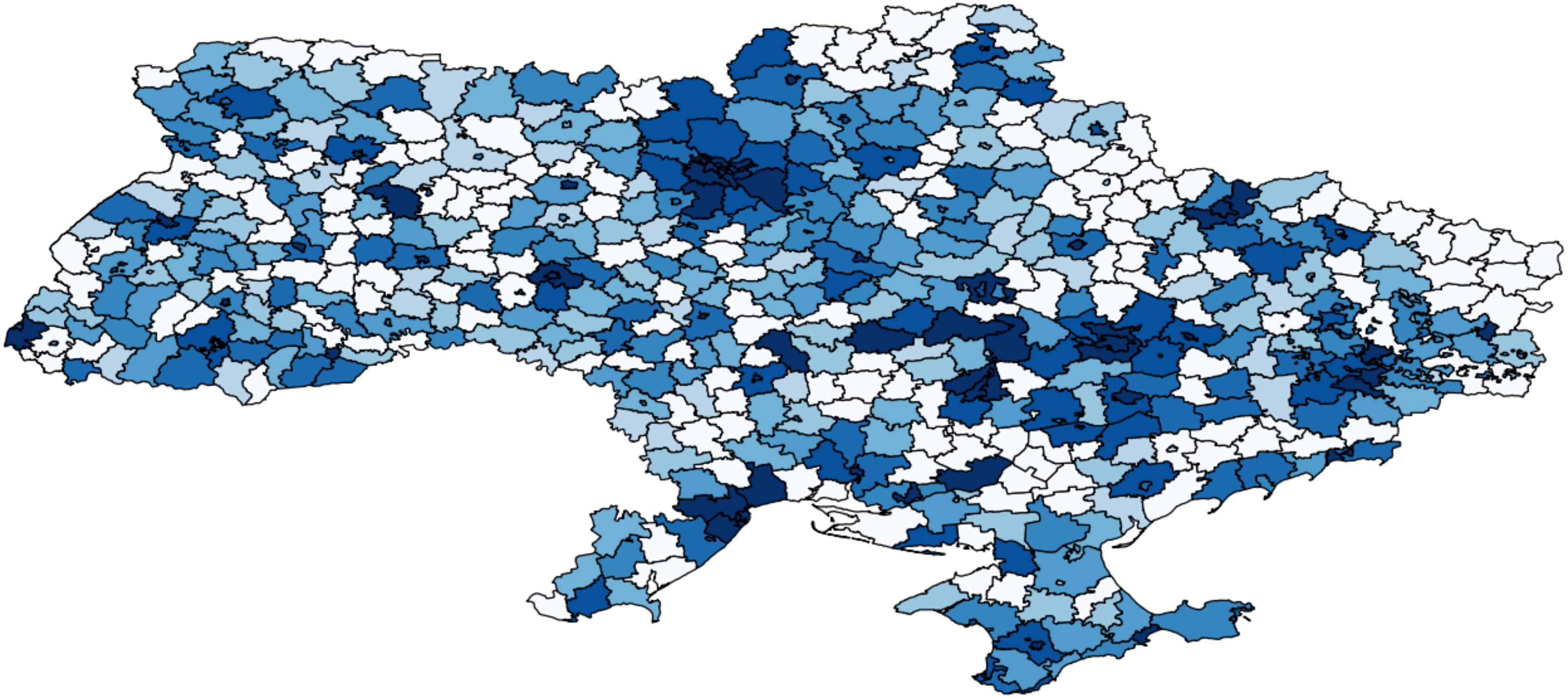
- Hand-coded from media sources
  - 26 days with size estimates (2K to 500K)
  - 73 days with protests but no size estimates
  - 95 days with no protest

Twitter in Central Kiev



Tweets (purple) and Violence (red) in Central Kiev, Feb 18-20, 2014

Empirical Testing



Tweets in Ukraine, Feb 18-20, 2014

How is Mass Protest Reflected in Social Media Activity?

- Overall twitter activity:
    - Increases nationwide
    - Decreases in capital
- Network-weighted twitter activity:
    - Decreases nationwide
    - Increases in capital
- Same center-periphery logic happens at regional capital level

Empirical Testing (Logit)

Variable	Sign	p	Sig
(Intercept)	-	<0.001	***
Core tweets	-	<0.001	***
Core network weighted	+	<0.001	***
Periphery tweets	+	<0.001	***
Periphery network weighted	-	<0.001	***
Time (# days)	-	.004	***

Logistic Regression of Occurrence of Protest on Social Media Variables (n=181)

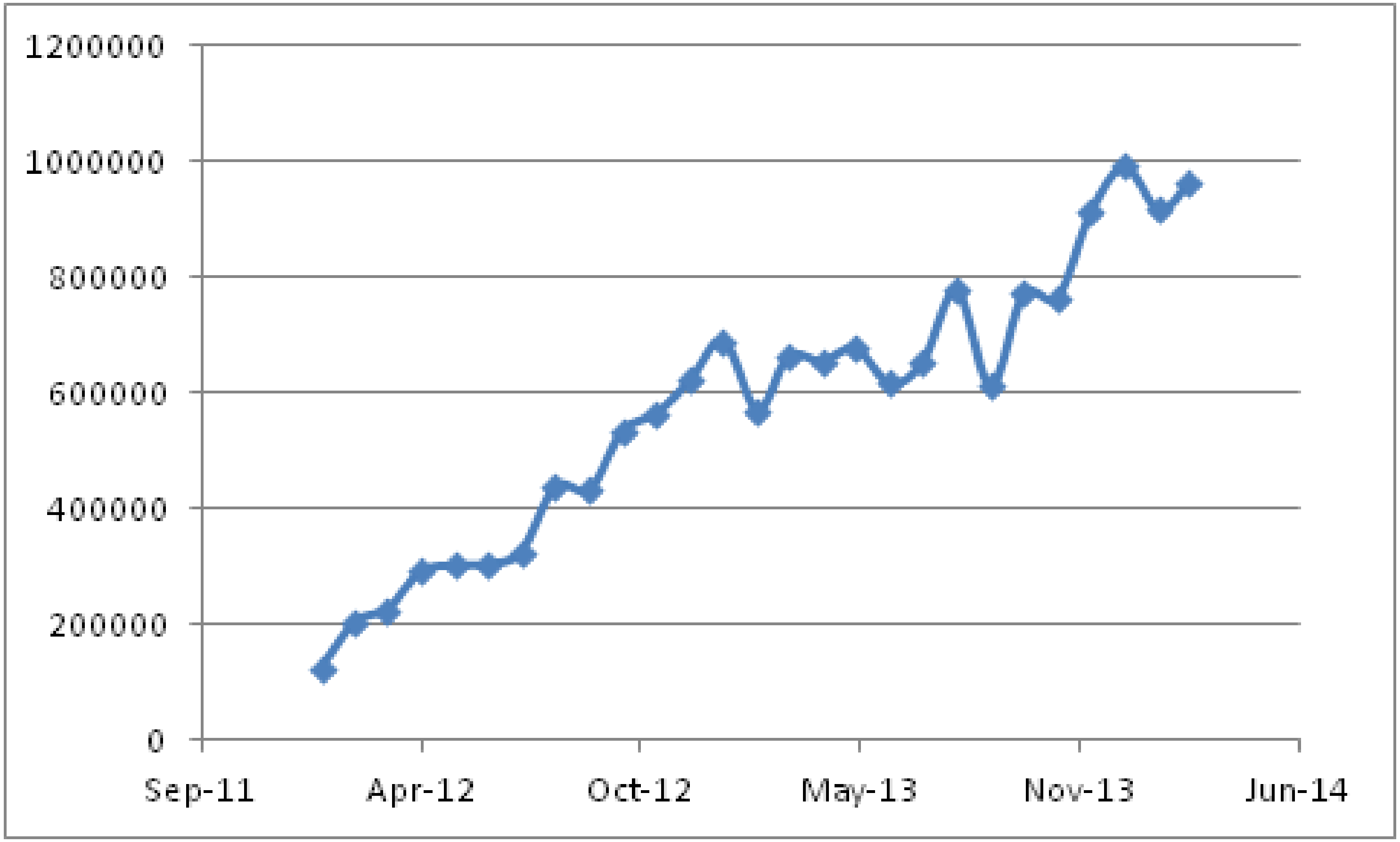
Empirical Testing (Logit)

	NP (pred)	P (pred)	Sum
No Protest (obs)	73	9	82
Protest (obs)	8	91	99
Sum	81	100	181

Classification Table of Logistic Regression

Wrapping Up

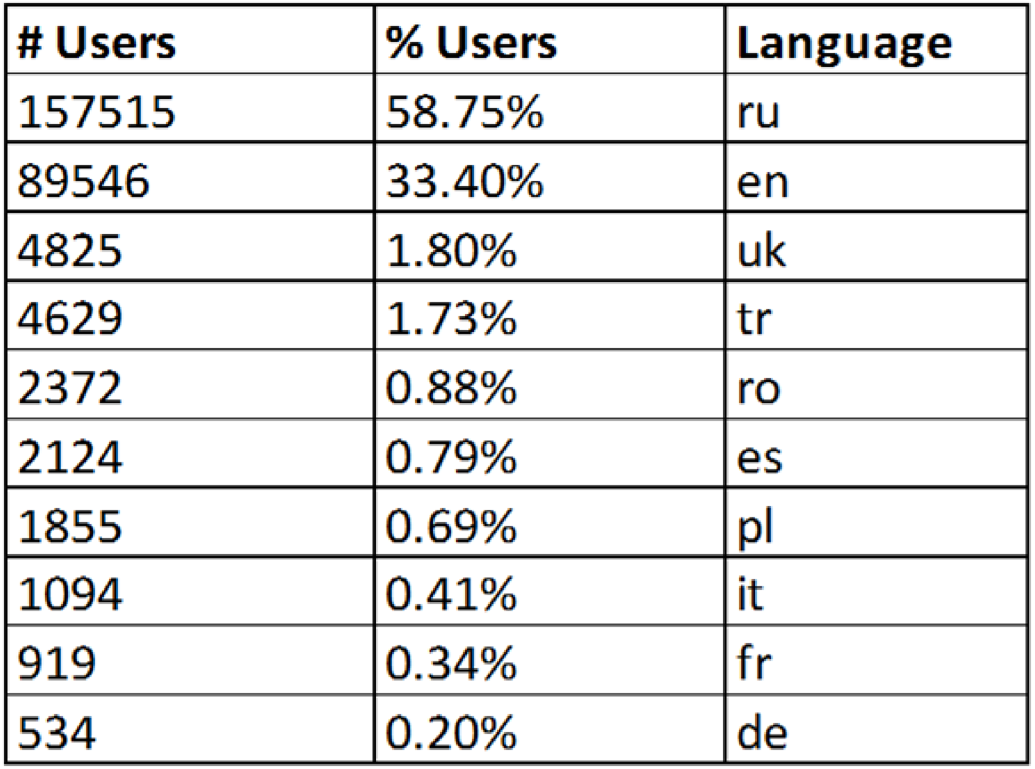
Twitter in Ukraine



Timezones

# Users	% Users	Time Zone
121132	45.18%	[Blank]
28734	10.72%	Kyiv
21512	8.02%	Moscow
15545	5.80%	Quito
14824	5.53%	Baghdad
13417	5.00%	Abu Dhabi
11830	4.41%	Athens
6576	2.45%	Bucharest
5231	1.95%	Greenland
3220	1.20%	Minsk

Language Settings



# Users	% Users	Language
157515	58.75%	ru
89546	33.40%	en
4825	1.80%	uk
4629	1.73%	tr
2372	0.88%	ro
2124	0.79%	es
1855	0.69%	pl
1094	0.41%	it
919	0.34%	fr
534	0.20%	de