MIS Quarterly

BIG DATA & ANALYTICS IN NETWORKED BUSINESS

CROWD-SQUARED: AMPLIFYING THE PREDICTIVE POWER OF SEARCH TREND DATA

Erik Brynjolfsson
Sloan School of Management, Massachusetts Institute of Technology,
Cambridge, MA 02142 U.S.A. {erikb@mit.edu}

Tomer Geva
School of Management, Tel-Aviv University,
Tel Aviv 6997801 ISRAEL {tgeva@tau.ac.il}

Shachar Reichman
School of Management, Tel-Aviv University, Tel Aviv 6997801 ISRAEL
and Sloan School of Management, Massachusetts Institute of Technology,
Cambridge, MA 02142 U.S.A. {shachar@mit.edu}

Appendix A

Crowd-Squared Search Terms Distribution

Table A1. Search Terms Generated by the Crowd

(a) Influenza – Crowd-Squared Search Terms and % of Turkers Mentioning Each Term

Search Term	% Mention	Search Term	% Mention	Search Term	% Mention	Search Term	% Mention
sick	68%	influenza	4%	achy	2%	fatigue	1%
fever	39%	coughing	4%	weak	2%	home	1%
cough	21%	germs	4%	bad	2%	hospital	1%
cold	16%	headache	4%	shots	2%	nyquil	1%
shot	13%	ache	4%	body aches	2%	season	1%
vomit	11%	runny nose	4%	flu shot	2%	tea	1%
ill	10%	pain	4%	sore throat	2%	under the weather	1%
sneeze	10%	aches	4%	soup	2%		
vaccine	10%	nausea	4%	tissue	2%		
virus	9%	bird	3%	vomiting	2%		
medicine	8%	miserable	3%	bug	2%		
contagious	8%	rest	3%	diarrhea	2%		
tired	8%	sore	3%	sweat	2%		

Table A1. Search Terms Generated by the Crowd (Continued)

(a) Influenza – Crowd-Squared Search Terms and % of Turkers Mentioning Each Term

Search Term	% Mention	Search Term	% Mention	Search Term	% Mention	Search Term	% Mention
sickness	8%	mucus	3%	nose	1%		
illness	7%	puke	3%	throwing up	1%		
doctor	7%	sneezing	3%	vaccination	1%		
bed	6%	swine	3%	nasty	1%		
snot	6%	congestion	3%	stuffy	1%		
chills	5%	death	3%	tissues	1%		
disease	5%	hot	2%	green	1%		
sleep	5%	stomach	2%	sad	1%		
gross	5%	winter	2%	achey	1%		

(b) Initial Claims for Unemployment Benefits – Crowd-Squared Search Terms and % of Turkers Mentioning Each Term

Search Term	% Mention	Search Term	% Mention	Search Term	% Mention	Search Term	% Mention
poor	23%	economy	3%	sadness	2%	stressful	1%
broke	20%	hungry	3%	search	2%	struggling	1%
jobless	15%	desperate	3%	unfortunate	2%	angry	1%
lazy	10%	hunger	3%	depressing	2%	applications	1%
money	10%	loss	3%	difficult	2%	bad	1%
sad	10%	out of work	3%	fear	2%	compensation	1%
no money	8%	work	3%	looking for work	2%	destitute	1%
poverty	8%	boredom	3%	rent	2%	job loss	1%
welfare	8%	family	3%	anger	1%	searching	1%
depression	7%	food	2%	check	1%	uncertainty	1%
job	7%	insurance	2%	helpless	1%	uneducated	1%
stress	6%	anxiety	2%	no work	1%	worry	1%
homeless	6%	failure	2%	scared	1%	assistance	1%
job search	5%	food stamps	2%	unlucky	1%	bad economy	1%
bills	5%	hardship	2%	boring	1%	despair	1%
no job	5%	help	2%	foreclosure	1%	frustrated	1%
laid off	4%	hopeless	2%	free time	1%	loser	1%
struggle	4%	job hunting	2%	frustration	1%	no insurance	1%
resume	4%	jobs	2%	hard times	1%	panic	1%
fired	4%	scary	2%	interviews	1%	rate	1%
depressed	3%	debt	2%	needy	1%	sucks	1%
bored	3%	government	2%	not working	1%	worthless	1%
benefits	3%	obama	2%	recession	1%		

Appendix B

Significance Values for Crowd-Squared Improvement Over Benchmark[1]

Our goal in the empirical evaluation section of this paper was to test whether our method, which is structured and transparent yet simple, achieves performance that is at least equivalent to the performance of existing benchmarks. We find that, in most cases, the crowd-squared method not only performs at the same level of existing benchmarks but actually obtains better results. In Table B1 below we report the results of significance tests regarding this comparison.

Table B1. Significance Test of Performance Improvement

Domain	Replicated Research and Evaluation Time Period	Benchmark Model	Significance for Crowd Squared Performance Improvement over Benchmark Model
Influenza Epidemics	Ginsberg et al. (2009), March 2007 – May 2008	Ginsberg et al.'s Model	
		Google Correlate	
		WordNet Lexicon	***
		Simple AR Benchmark	***
	Ginsberg et al. (2009), October 2012 – September 2013 (Period as in Butler 2013)	Ginsberg et al.'s Model	*
		Google Correlate	*
		WordNet Lexicon	***
		Simple AR Benchmark	***
	Lazer et al. (2014)	Google Flu Trends	***
		Lazer et al.'s Model	*
Initial Claims for Unemployment Benefits	Choi and Varian (2012)	AR Model	
		Choi and Varian's Model	n/a[a]
		Google Correlate	*
		WordNet Lexicon	***

**p* value < 0.1   ***p* value < 0.05   ****p* value < 0.01

[a] We do not include results pertaining to Choi and Varian's (2012) model. This is due to recent changes in Google Trends categories that prohibit using the same categories and reconstructing the weekly level predictions and prediction errors of Choi and Varian's model. Nevertheless, we note that our results were significantly better than the Google Correlate model's results over the same data, which, in turn, outperforms Choi and Varian's reported results

[1] The different studies that we replicated used different performance measures and required different methods for significance value calculations. We used bootstrap p-values for significance testing when replicating the study by Ginsberg et al. (2009), which used correlation with CDC-reported ILI as a performance measure. We used the Diebold-Mariano test, which was used in Lazer et al. (2014), to calculate significance values for the MAE performance measure reported in Lazer et al. and in Choi and Varian (2012).

Appendix C

Participant Demographics

Table C1. Reported Demographics for Crowd Squared Participants and Data Collection Costs

	Influenza Epidemics (N = 535)	Initial Claims for Unemployment Benefits (N = 545)
Age	Avg. 29.1 (std. 9.12)	Avg. 33 (std. 11)
Gender	40% female	58.5% female
Education Level		
• Bachelor degree	210	206
• Master degree	39	56
• Some College	212	209
• Professional degree	10	12
• High School	53	56
• Doctorate degree	11	6
Number of Participants' States	48	47
Mechanical Turk Cost	$35.31	$32.11

Appendix D

Correlation as a Performance Measure

As detailed in the main body of the paper, we used an experimental design geared to provide a fair comparison with previous studies. This requires that when comparing the crowd-squared method to an alternative methodology used in a prior study, we use the same goal and performance measures adopted in that study. Specifically, Ginsberg et al. (2009) aimed to obtain the best correlation, whereas Lazer et al. (2014) and Choi and Varian (2012) sought to minimize MAE. Therefore, in different comparisons, our method was evaluated on the basis of different performance measures.

Nevertheless, for robustness we present in Tables D1 and D2 correlation results for models originally set to optimize MAE. As shown in these tables, our method obtains comparable or superior results, in terms of correlation, compared to the benchmark studies and models, even though the goal we set was to improve MAE. We note, however, that performance improvement using correlation values for the models set to improve MAE was (expectedly) smaller in scale.

Table D1. Correlation Results with CDC-Reported ILI for Prediction Models Using Different Data Selection Methods (Comparison to Lazer et al. 2014)

Crowd-Squared	Lazer et al.	Google Flu Trends
0.96722	0.95608	0.86779

Table D2. Correlation Results with Initial Claims for Unemployment Benefits Using Different Data Selection Methods

Crowd-Squared	AR (1) Model	Choi and Varian	Google Correlate	WordNet Lexicon
0.98137	0.98132	n/a[a]	0.98074	0.98035

[a]We do not include results pertaining to Choi and Varian's model. This is due to recent changes in Google Trends categories that prohibit using the same categories and reconstructing the weekly-level predictions and prediction errors of Choi and Varian's model. Nevertheless, we note that our results were significantly better than the Google Correlate model's results over the same data, which, in turn, outperforms Choi and Varian's reported results

Appendix E

Additional Analysis 2014–2015

In the main body of the paper, we provide a comparison of our data selection method performance with data selection methods used in previous studies while using the same time periods reported in these studies. For robustness, we evaluate whether the crowd-squared method could provide comparable or better results to the above benchmarks studies and models on recent data. For this purpose we evaluate our method performance over a full year of out-of-sample data (May 1, 2014–April 30, 2015) immediately following the month in which we ran the online word-association task (April 2014). The results are detailed in Tables E1 and E2 10 and show that in accordance with the previous findings, our method obtain comparable or better results to the benchmarks.

Table E1. Correlation Results with CDC-Reported ILI for Prediction Models Using Different Data Selection Methods[a]

Crowd-Squared	Google Flu Trends	Google Correlate	WordNet Lexicon	Simple Benchmark
0.982	0.985	0.976	0.972	0.937

[a]Recently Google reported about changes to their flu trend system which now also uses lagged CDC data as predictors. Unfortunately, at the current time Google has not yet provided details about the specifics of their method. For the sake of comparison with the Google flu trend system we added to our model only a single lag of CDC data (lag t-2) and weekly dummy variables (as in the Lazer et al. reference that Google report provided). See also http://googleresearch.blogspot.co.il/2014/10/google-flu-trends-gets-brand-new-engine.html, accessed July 2015.

Table E2. MAE Results for Predicting Initial Claims for Unemployment Benefits Using Different Data Selection Methods

Crowd-Squared	AR (1) Model	Choi and Varian (2012)	Google Correlate	WordNet Lexicon
3.37%	3.44%	n/a[a]	3.75%	3.47%

[a]We do not include results pertaining to Choi and Varian's model. This is due to recent changes in Google Trends categories that prohibit using the same categories and reconstructing the weekly-level predictions and prediction errors of Choi and Varian's model.

Appendix F

Cost and Time Estimation

	Comprehensive Scan[a]		Crowd- Squared	Prior Knowledge and Intuition[b]
	1 billion queries	50 million queries		
Data download (computer hours)	~55,000	~2,800	0.007 computer hours (~100 queries)	0.007 computer hours (up to 100 queries)
Data storage	~ 20 TB	~ 1TB	~ 2 MB	~ 2MB
Model analysis and predictions(computer hours)	~20 hours	~20 hours	< 1 hour	< 1 hour
Payments to participants (through AMT)	—	—	500 participants × $0.06 per task = $30 + 10% AMT platform cost = $33	—
Data Scientists	2days	2 days	2days	2 days

[a]Based on the analysis process reported in Ginsberg et al. (2009).
[b]Based on the analysis process reported in Choi and Varian et al. (2012).

Cost Estimation Assumptions

- Comprehensive scan includes two options:
  - Downloading data for 1 billion queries and then finding the most popular 50 million search queries, assuming there is no prior information about the most popular search terms.
  - Downloading data on the most popular 50 million queries, assuming that the search engine publishes information on the popularity of search terms.
- Average download time for query trend data: 0.2 second.
- Average file size for data for a single query trend: 20 KB.
- 500 participants in the crowd-squared tasks.
- Average payment to participant: $0.06.
- Expert (data scientists) time to perform the complete analysis in any method: 2 days.

References

Choi, H., and Varian, H. 2012. "Predicting the Present with Google Trends," *Economic Record* (88:s1), pp. 2-9.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., and Brilliant, L. 2009. "Detecting Influenza Epidemics Using Search Engine Query Data," *Nature* (457:7232), pp. 1012-1014.

Lazer , D., Kennedy, R., King, G., and Vespignani, A. 2014. "The Parable of Google Flu: Traps in Big Data Analysis," *Science* (343:6176), pp. 1203-1205