

Analyzing and Detecting Opinion Spam on a Largescale Dataset via Temporal and Spatial Patterns Huayi Li[†], Zhiyuan Chen[†], Arjun Mukherjee [‡], Bing Liu[†] and Jidong Shao^{*}

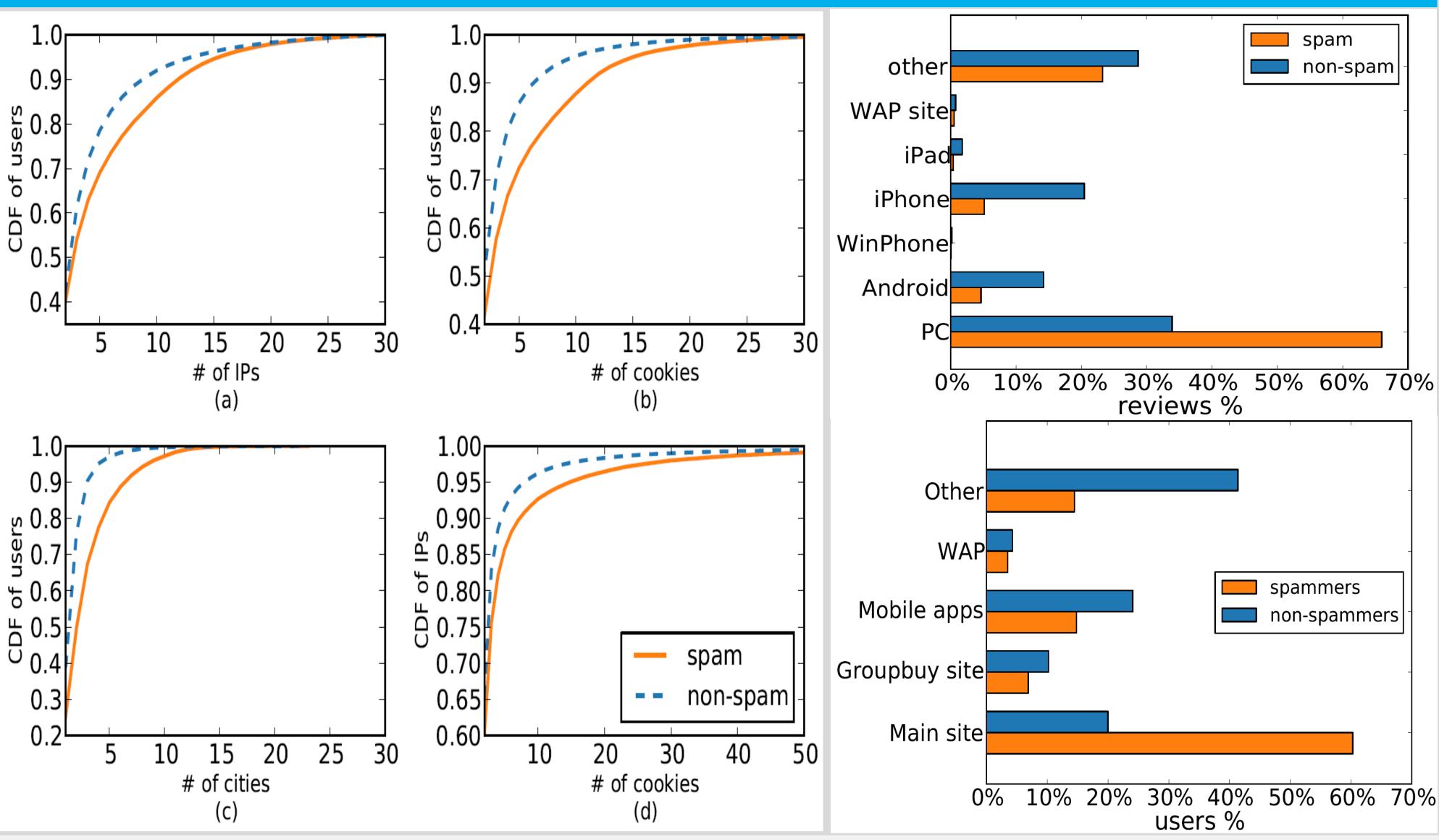
† Department of Computer Science, University of Illinois at Chicago, IL, USA ‡ Department of Computer Science, University of Houston, TX, USA * Dianping Inc., Shanghai, China

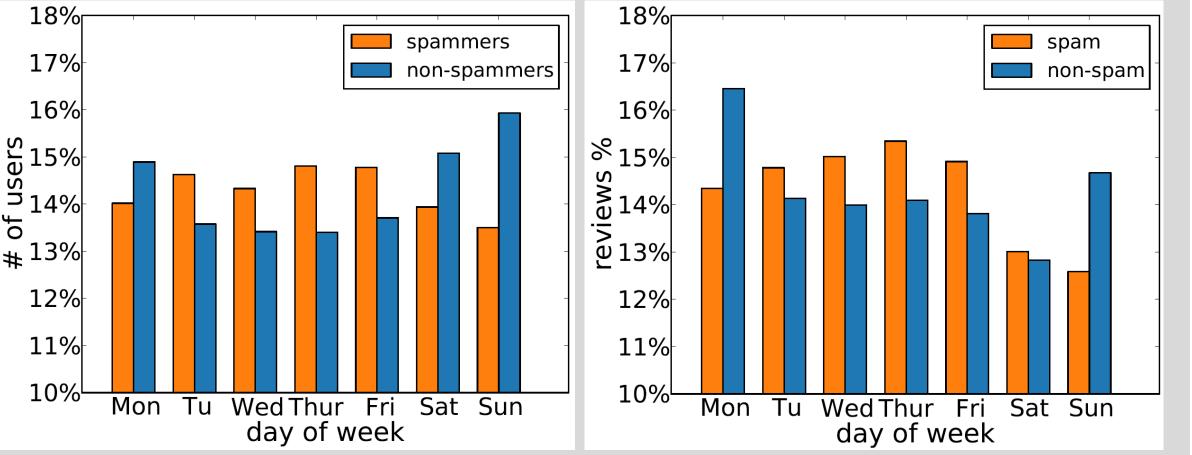
Summary

- Data Volume: Over 6 million reviews. no existing study has been performed on such a large scale.
- Data Richness: Rich attributes, including users' IP addresses, users' profile and so on.
- **Feature Novelty:** The first to give comprehensive insights of temporal and spatial features. Those features help build more accurate classification models.

Temporal Patterns

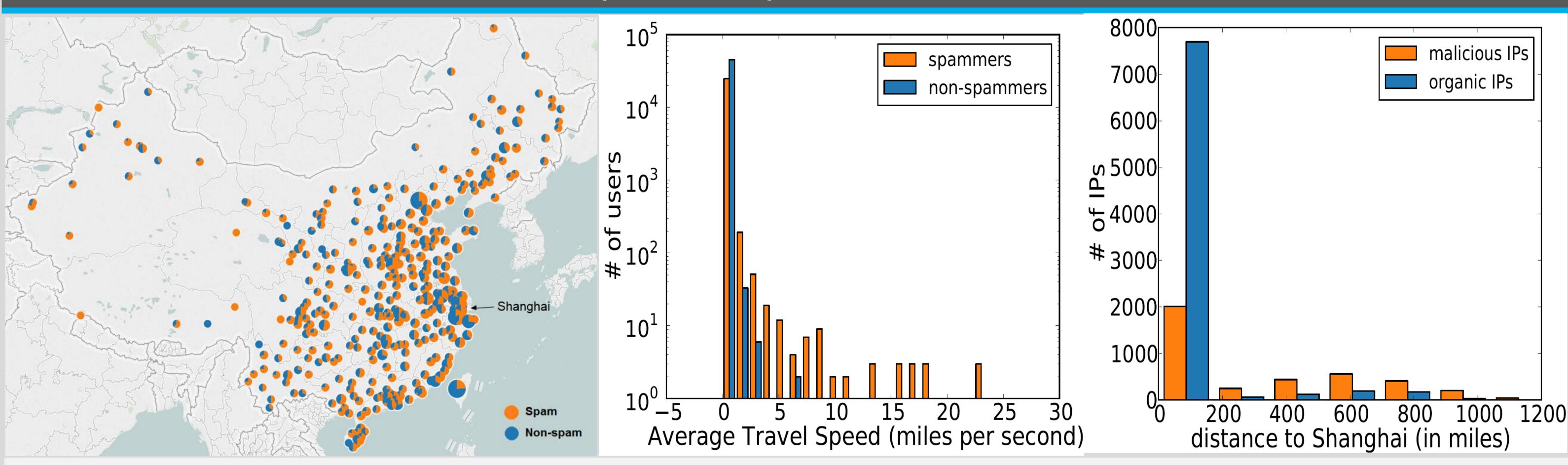






- Authentic reviews are much more than fake reviews on Mondays and Sundays because they are based on real experience of dinners at weekends
- Similarly, more non-spammers are registered at weekends and Mondays
- Registration and composing reviews in main site is the fastest way, so spammers show a
 preference in registering and posting reviews on the main site
- Spammers switch IP addresses/cities and even browser cookies more often than ordinary/genuine users

Spatio-temporal Patterns



People in large cities (a few big pie charts) are dominated by non-spammers

The further the cities are from Shanghai, the higher the ratios of spammers

Users who frequently and randomly "move" all over China with unusual speeds are likely to be spammers

Experiments Datasets Features Restaurant reviews in Shanghai, China

Method	Accuracy	Precision	Recall	F1	If a user is registered in main site	from November 1st, 2011 to April 18th, 2014			
Unigram and Bigram	0.68	0.71	0.63	0.67	If a user is registered between Tu. and Thur.	#Reviews	#users	#IPs	#restaurants
					Distance from the registration city to Shanghai				
					Average Travel Speed	6,126,113	1,074,604	1,331,471	108,787
Behavioral features	0.74	0.71	0.78	0.73	% of reviews posted at weekends				
					% of reviews posted through PC	Acknowledge			
Proposed Features	0.84	0.81	0.86	0.83	Average distance to Shanghai				
					Average absolute rating deviation	We would like to thank the spam detection team in Dianping for generously sharing the review			
					# of unique IPs used by a user	dataset. This paper is made possible through the			
Combined	0.85	0.83	0.87	0.85	# of unique cookies by a user	help and support from engineers and scientists in Dianping who provided valuable suggestions and indispensable efforts in evaluation			
					# of unique cities where a user write reviews				