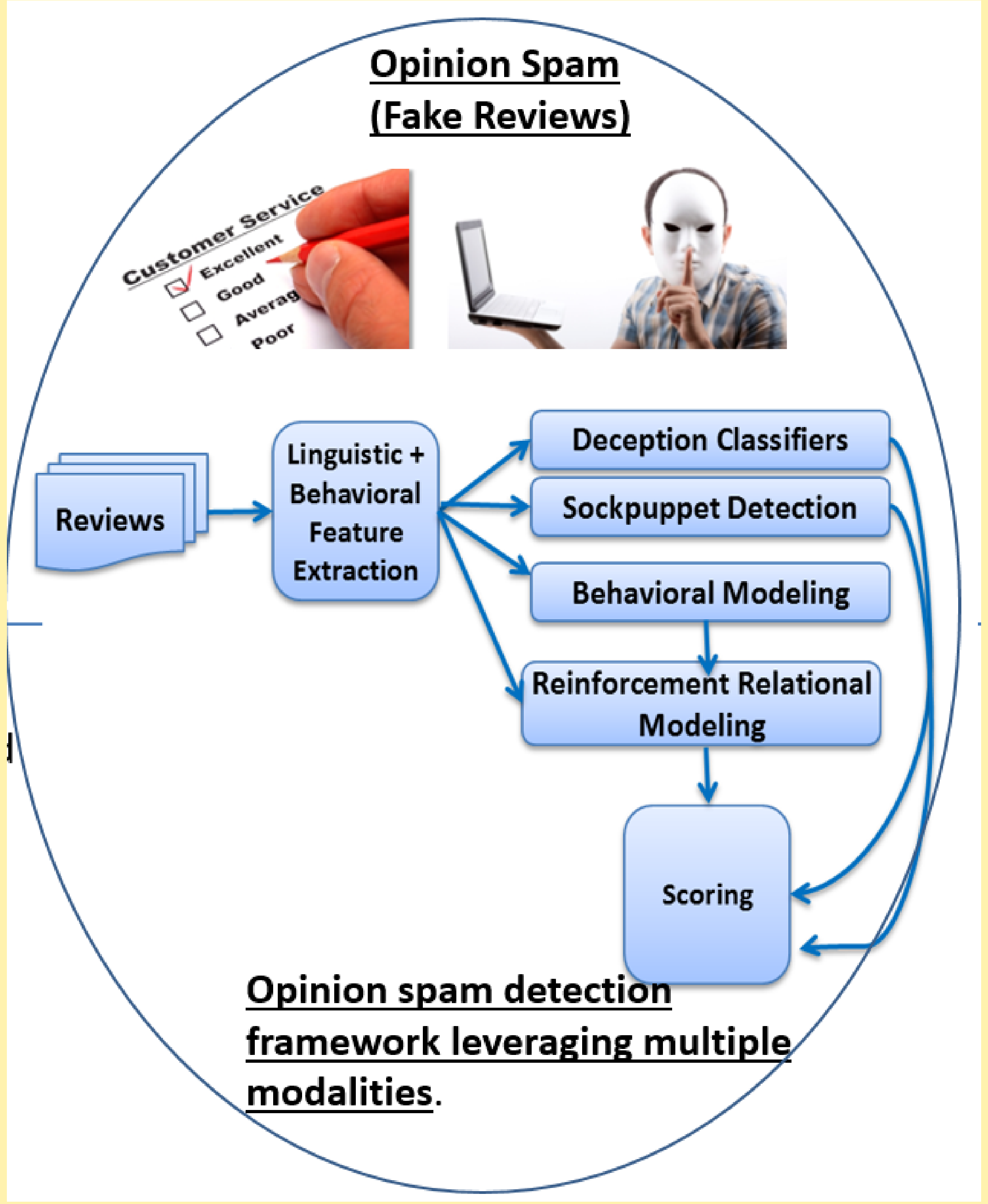
Statistical Models for Opinion Spam Detection Leveraging Linguistic and Behavioral Cues

PI: Arjun Mukherjee SP: Thamar Solorio

This project aims to detect fake reviews (opinion spam) using a synergistic combination of linguistic, behavioral and statistical methods.

IM: Focus on learning and evaluation on unlabeled data. Strategies can be applied to sockpuppets, viral hoaxes, forensic linguistics, and other abuse

BI: Improve trustworthiness of web content and reduce social implausibility. Characterizes several behavioral modalities of deception posing a marketing, consumer, and economic risk



Opinion spam detection framework leveraging multiple modalities.

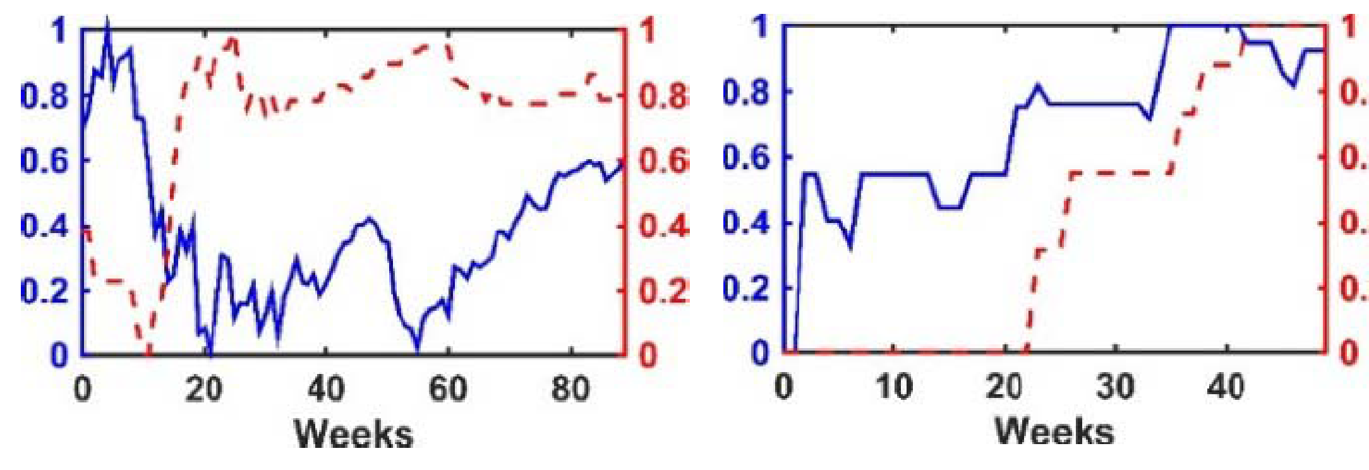
Multi-pronged Approach via Linguistics, Behavioral and Statistical Modeling

- Leverage small scale domain expert labeled data with large crowd sourced fake reviews to bootstrap fake review detection (e.g., employ corrective learning and ensemble methods)
- Model spamicity as latent with observed behavioral gfootprints

- Exploit latent reinforcement relations between (1) fraudsters, (2) fake reviews, (3) deceptive language, and (4) fraud behavior. (e.g., knowing fraudsters can help discover other fraudsters who reviewed the same entity or exhibit similar linguistic/behaviors
- Sockpuppet modeling using transductive learning in orthogonal language spaces

Progress so far

- Explored various temporal spamming patterns and policies spammers generally employ and showed that temporal dynamics are very helpful in predicting spam [Kc and Mukherjee WWW 2016]



a) Truthful Like Rating b) Truthful Dislike Rating

Figure 2. Buffered Spamming - Time series of truthful like (> 3star) and dislike (<=3 star) ratings (solid blue) vs. deceptive like rating (dashed red) for different representative restaurants. Representative restaurants refers to the ones where the behavior was most prominent.

More Aspects of Progress

- A spy induction framework has been explored for detecting sockpuppets. The problem was formulated as a author verification task and samples of reviews written by a spam reviewer were used to find other potential reviews written under different alias.

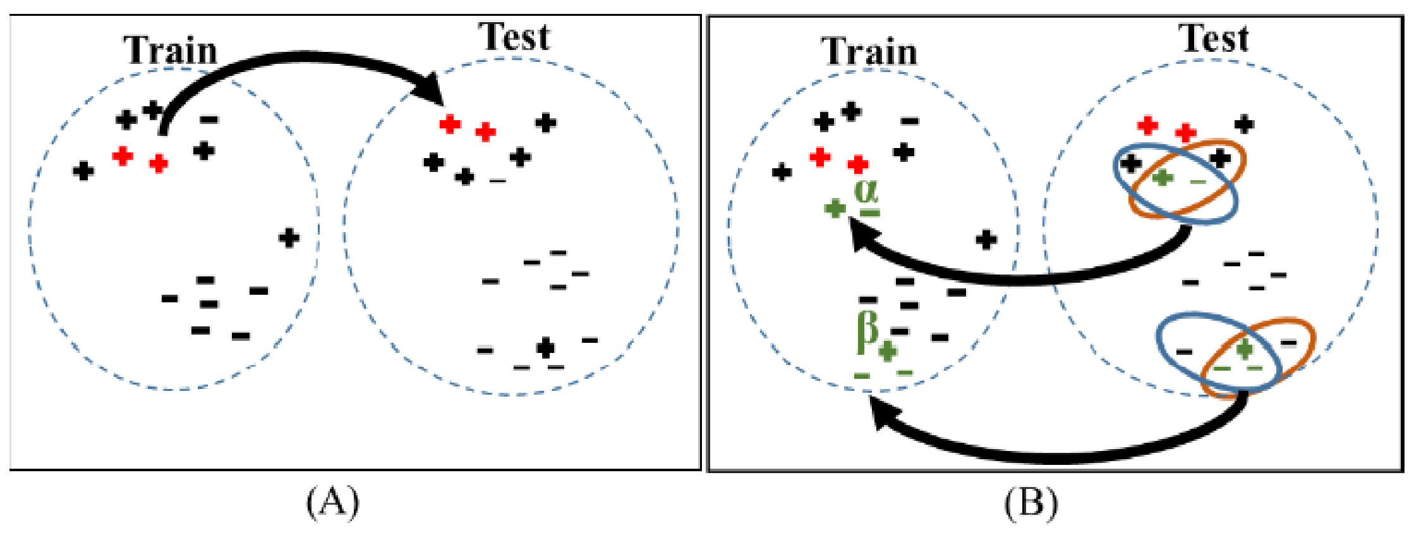
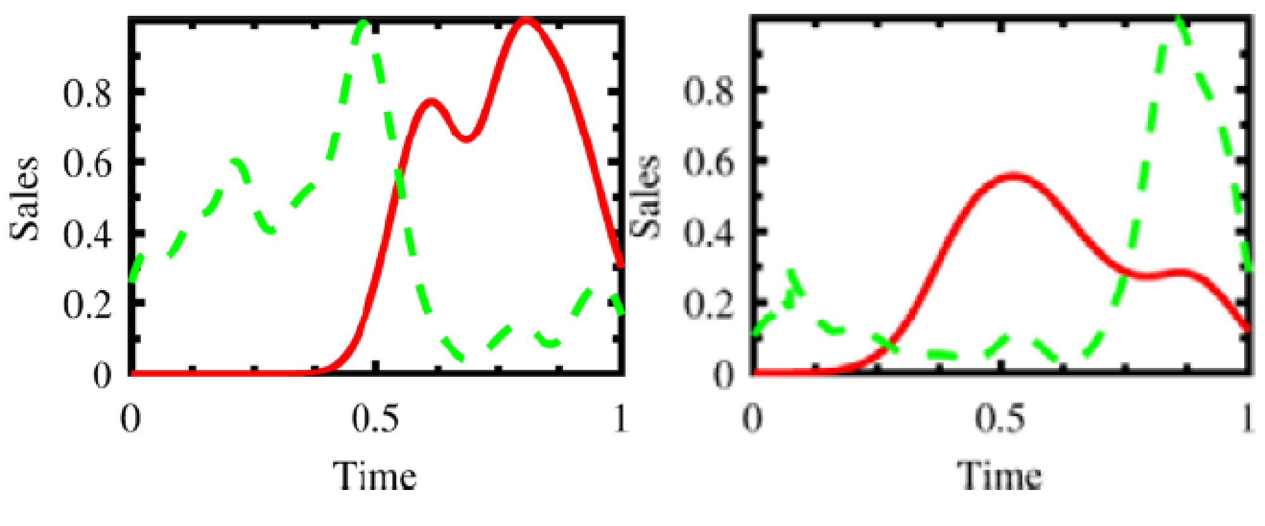


Figure 7: Spy Induction : (A)Spies (Red plus signs) selected based on +ve class centrality being put to the unlabeled test set. (B) Common nearest and farthest neighbors (Green plus and minus signs) across different spies' neighborhood shown by oval boundaries found in unlabeled test set are being put back in the train set.

Still More Aspects of Progress and Future Plans

Explore the phenomena of market competition to yield evidences as competition often triggers review spam injection.

When does a takeover happen? When does a recovery happen? By what volume do sales takeover?



a) Death *b) Survival*

Figure 4. Scaled normalized sales time-series depicting competition between a previously Leading product (in dashed green line) by a Competitor (solid red line).

Interested in meeting the PIs? Attach post-it note below!

NSF Secure and Trustworthy Cyberspace Principal Investigator Meeting
Jan 9-11, 2017, Arlington, VA