CS47300: Web Information Search and Management

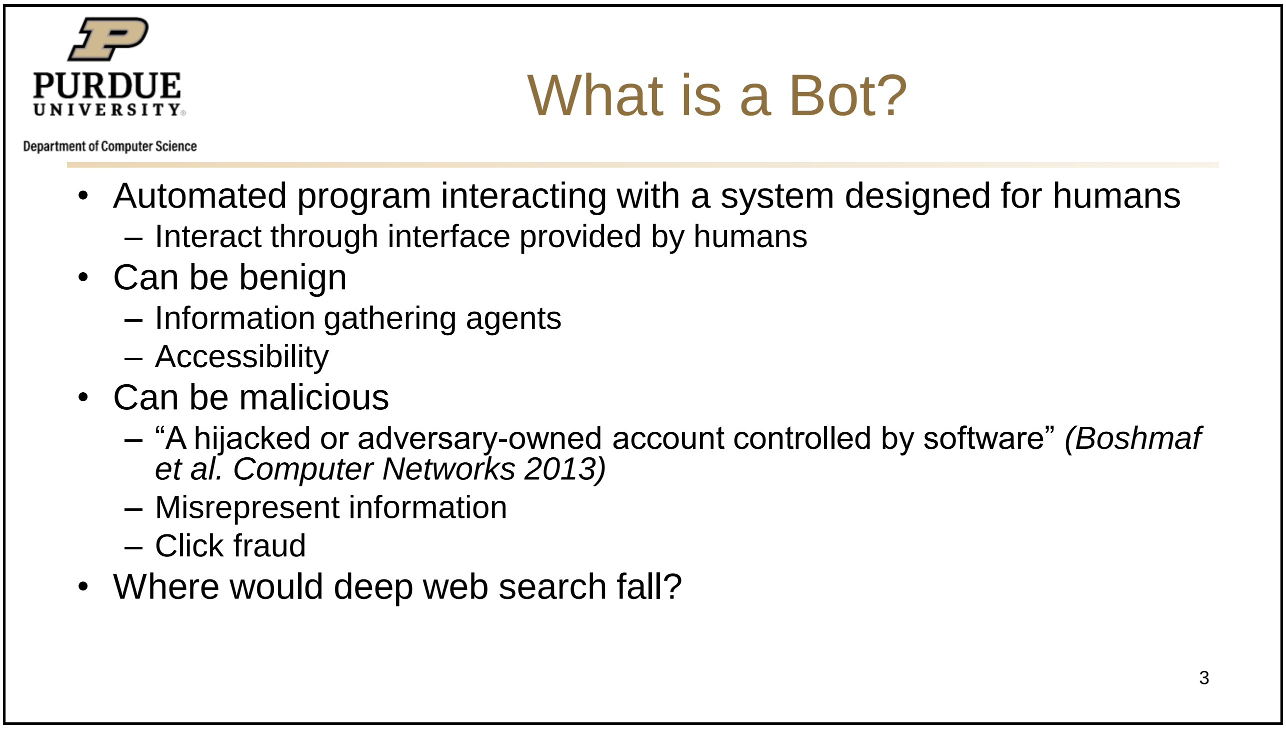
Bot Detection

Prof. Chris Clifton

23 November 2020

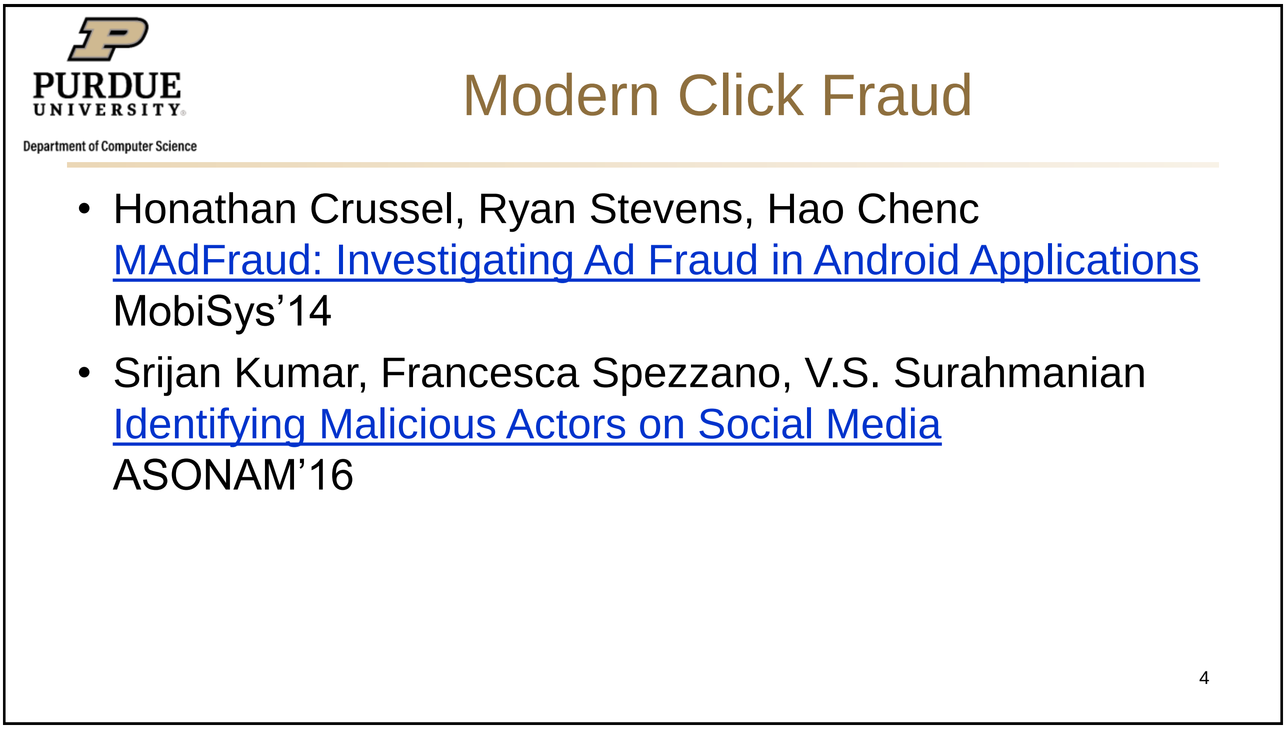
Bad Actors in IR

- Search Engine Optimization
  - Can be good
  - Can be bad
- Fake News
  - And other fake content
- Bots acting as human
  - Fake content
  - Fake behavior
- Others?



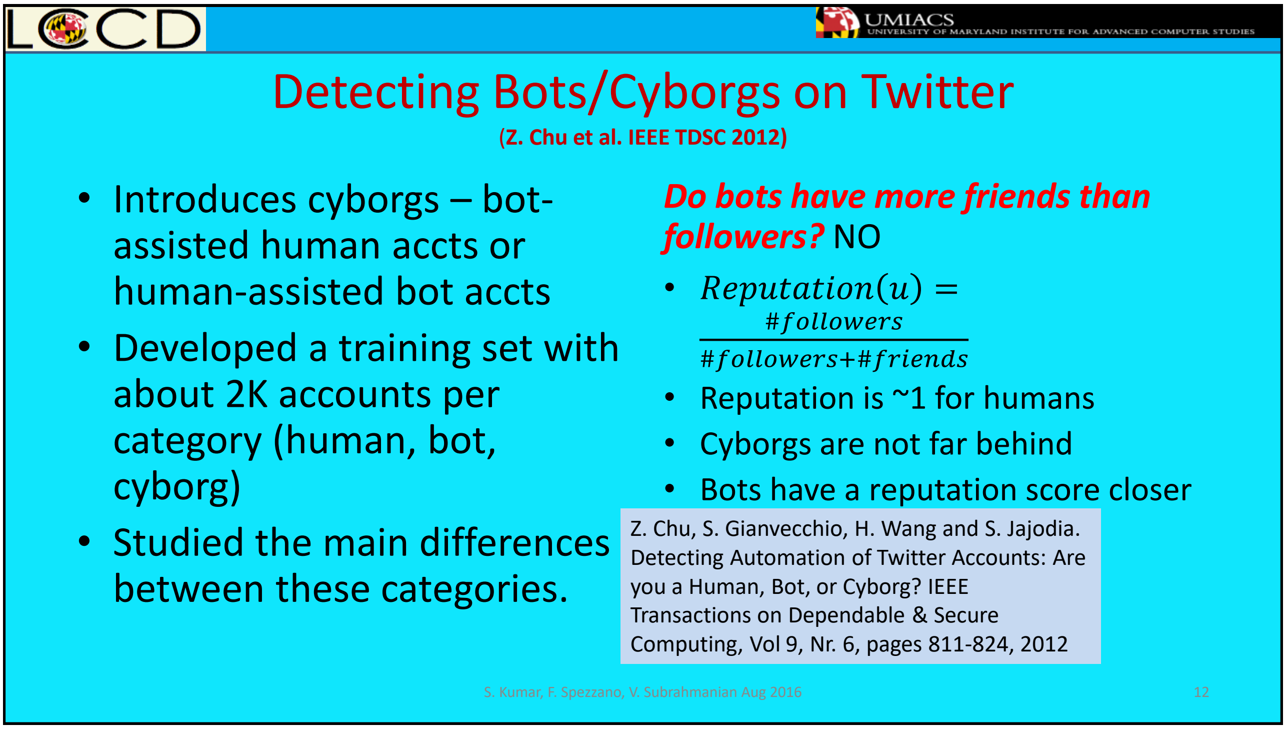
What is a Bot?

- Automated program interacting with a system designed for humans
  - Interact through interface provided by humans
- Can be benign
  - Information gathering agents
  - Accessibility
- Can be malicious
  - “A hijacked or adversary-owned account controlled by software” *(Boshmaf et al. Computer Networks 2013)*
  - Misrepresent information
  - Click fraud
- Where would deep web search fall?



Modern Click Fraud

- Honathan Crussel, Ryan Stevens, Hao Chenc
  MAdFraud: Investigating Ad Fraud in Android Applications
  MobiSys’14
- Srijan Kumar, Francesca Spezzano, V.S. Surahmanian
  Identifying Malicious Actors on Social Media
  ASONAM’16



Detecting Bots/Cyborgs on Twitter

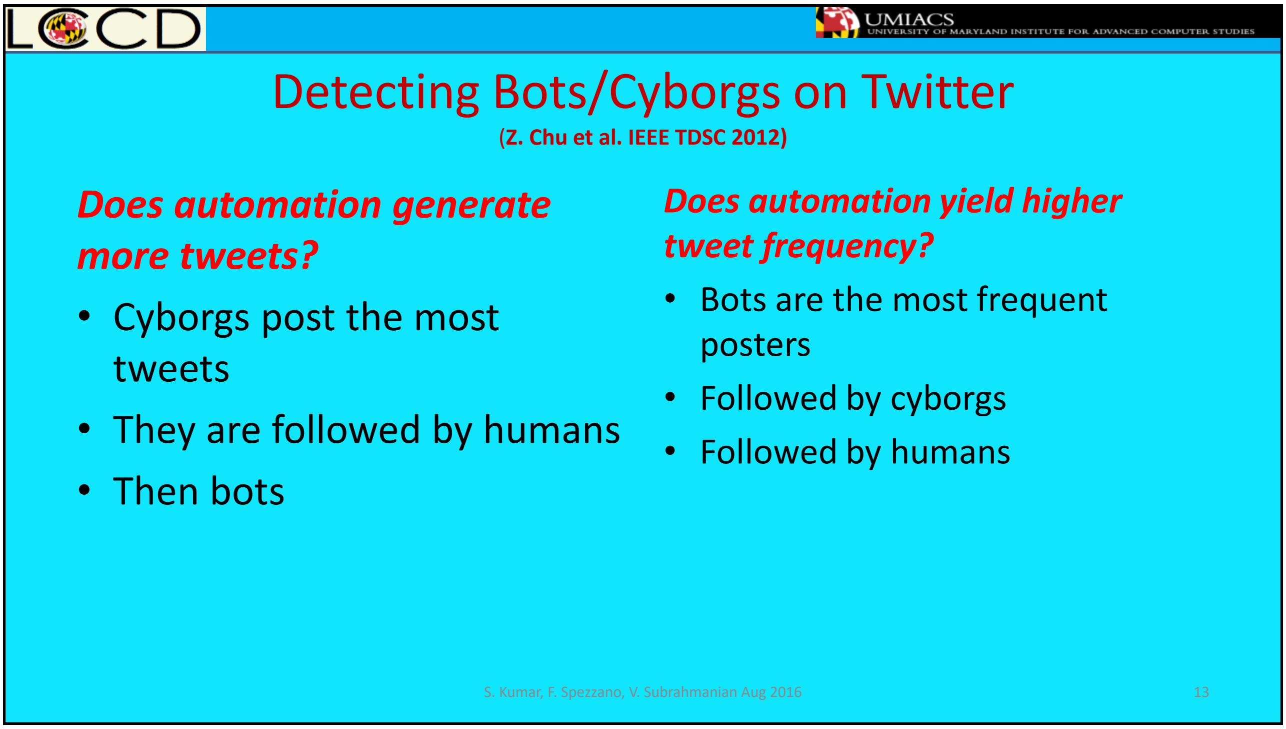
(Z. Chu et al. IEEE TDSC 2012)

- Introduces cyborgs – bot-assisted human accts or human-assisted bot accts
- Developed a training set with about 2K accounts per category (human, bot, cyborg)
- Studied the main differences between these categories.

Do bots have more friends than followers? NO

- $Reputation(u) = \frac{\#followers}{\#followers+\#friends}$
- Reputation is ~1 for humans
- Cyborgs are not far behind
- Bots have a reputation score closer

Z. Chu, S. Gianvecchio, H. Wang and S. Jajodia. Detecting Automation of Twitter Accounts: Are you a Human, Bot, or Cyborg? IEEE Transactions on Dependable & Secure Computing, Vol 9, Nr. 6, pages 811-824, 2012



Detecting Bots/Cyborgs on Twitter

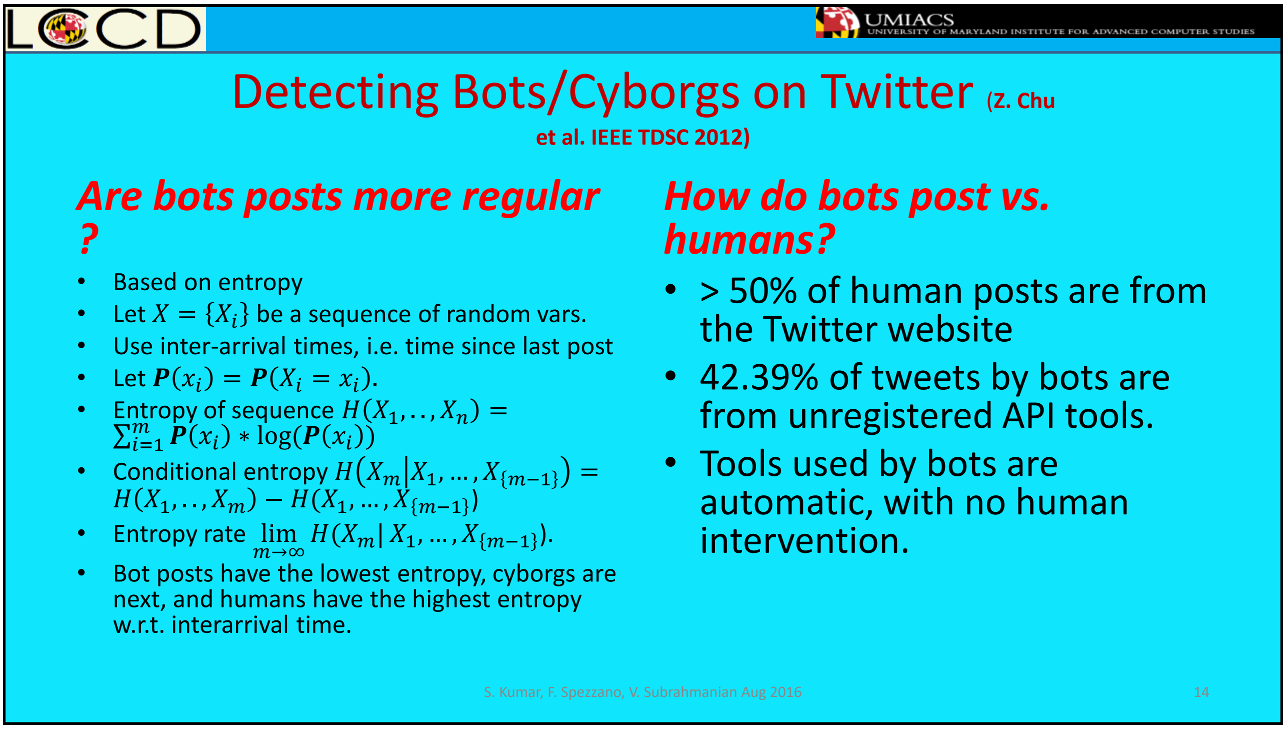
(Z. Chu et al. IEEE TDSC 2012)

Does automation generate more tweets?

- Cyborgs post the most tweets
- They are followed by humans
- Then bots

Does automation yield higher tweet frequency?

- Bots are the most frequent posters
- Followed by cyborgs
- Followed by humans



UMIACS UNIVERSITY OF MARYLAND INSTITUTE FOR ADVANCED COMPUTER STUDIES

Detecting Bots/Cyborgs on Twitter (Z. Chu et al. IEEE TDSC 2012)

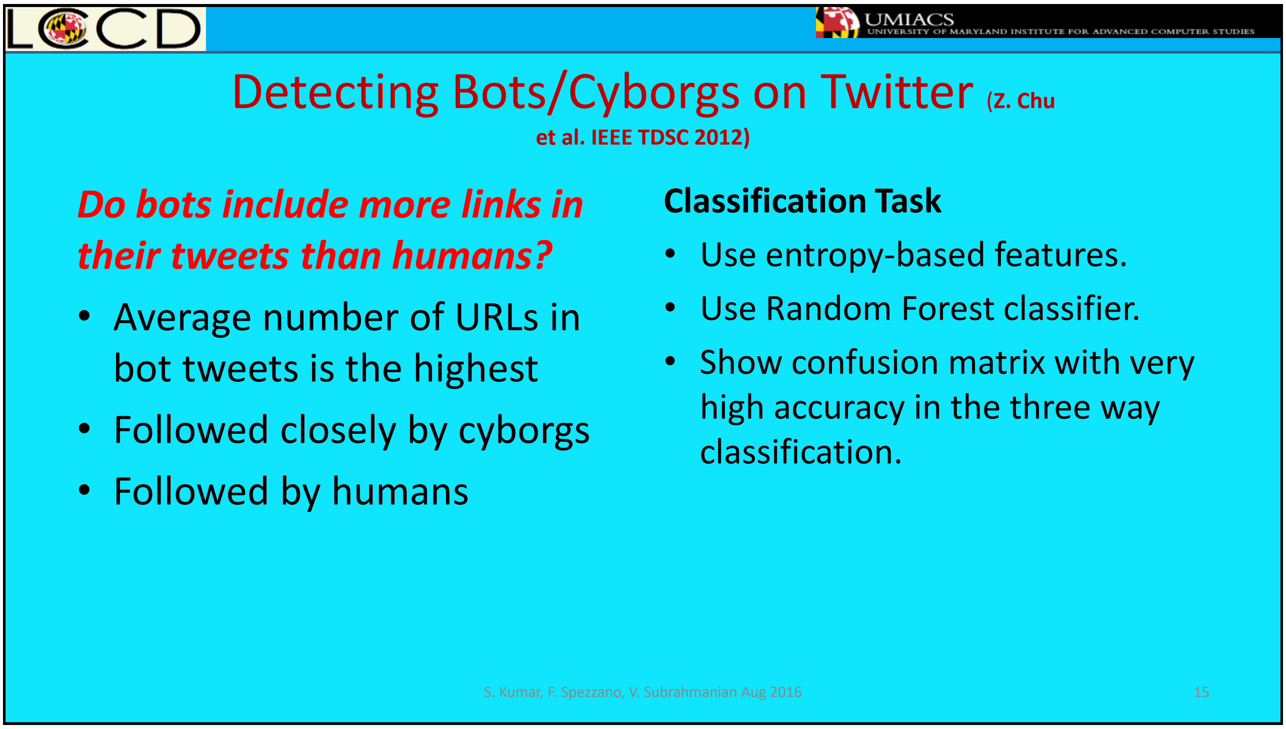
Are bots posts more regular ?

- Based on entropy
- Let $X = \{X_i\}$ be a sequence of random vars.
- Use inter-arrival times, i.e. time since last post
- Let $\boldsymbol{P}(x_i) = \boldsymbol{P}(X_i = x_i)$.
- Entropy of sequence $H(X_1, .., X_n) = \sum_{i=1}^{m} \boldsymbol{P}(x_i) * \log(\boldsymbol{P}(x_i))$
- Conditional entropy $H(X_m | X_1, \dots, X_{\{m-1\}}) = H(X_1, .., X_m) - H(X_1, \dots, X_{\{m-1\}})$
- Entropy rate $\lim_{m \to \infty} H(X_m | X_1, \dots, X_{\{m-1\}})$.
- Bot posts have the lowest entropy, cyborgs are next, and humans have the highest entropy w.r.t. interarrival time.

How do bots post vs. humans?

- > 50% of human posts are from the Twitter website
- 42.39% of tweets by bots are from unregistered API tools.
- Tools used by bots are automatic, with no human intervention.

S. Kumar, F. Spezzano, V. Subrahmanian Aug 2016 14



UMIACS UNIVERSITY OF MARYLAND INSTITUTE FOR ADVANCED COMPUTER STUDIES

Detecting Bots/Cyborgs on Twitter (Z. Chu et al. IEEE TDSC 2012)

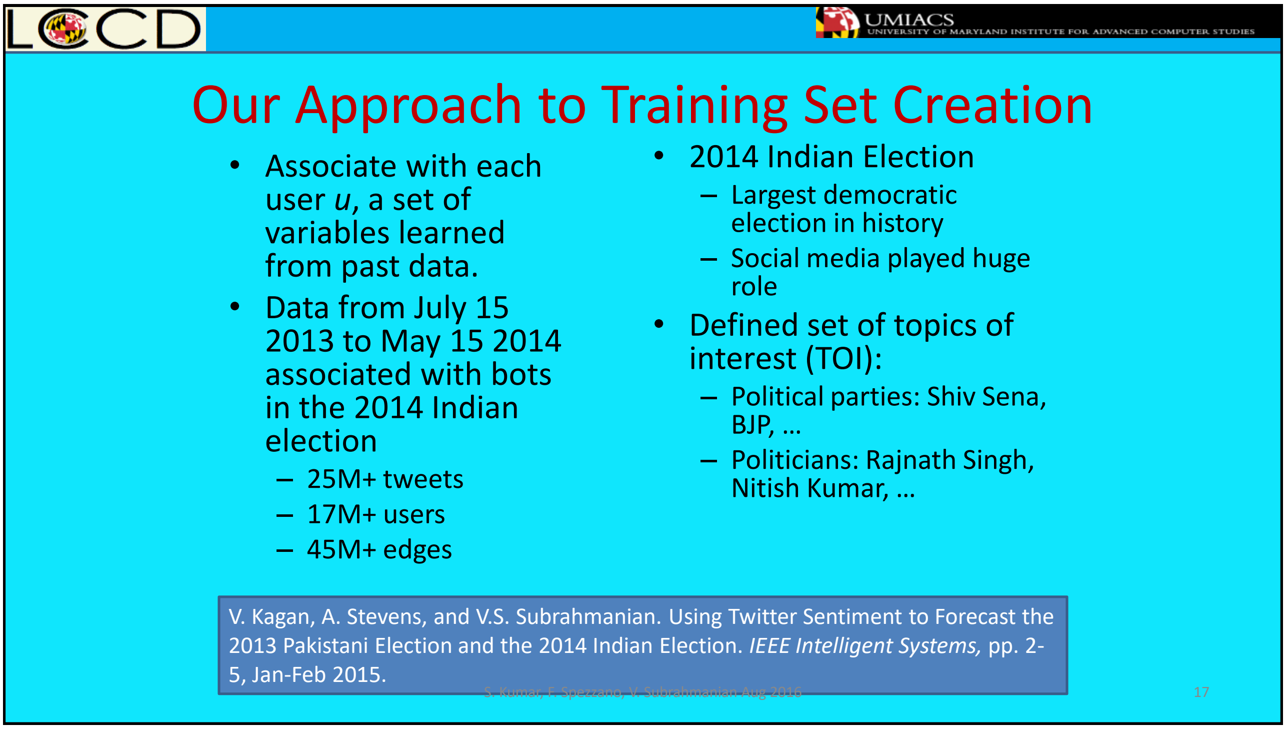
Do bots include more links in their tweets than humans?

- Average number of URLs in bot tweets is the highest
- Followed closely by cyborgs
- Followed by humans

Classification Task

- Use entropy-based features.
- Use Random Forest classifier.
- Show confusion matrix with very high accuracy in the three way classification.

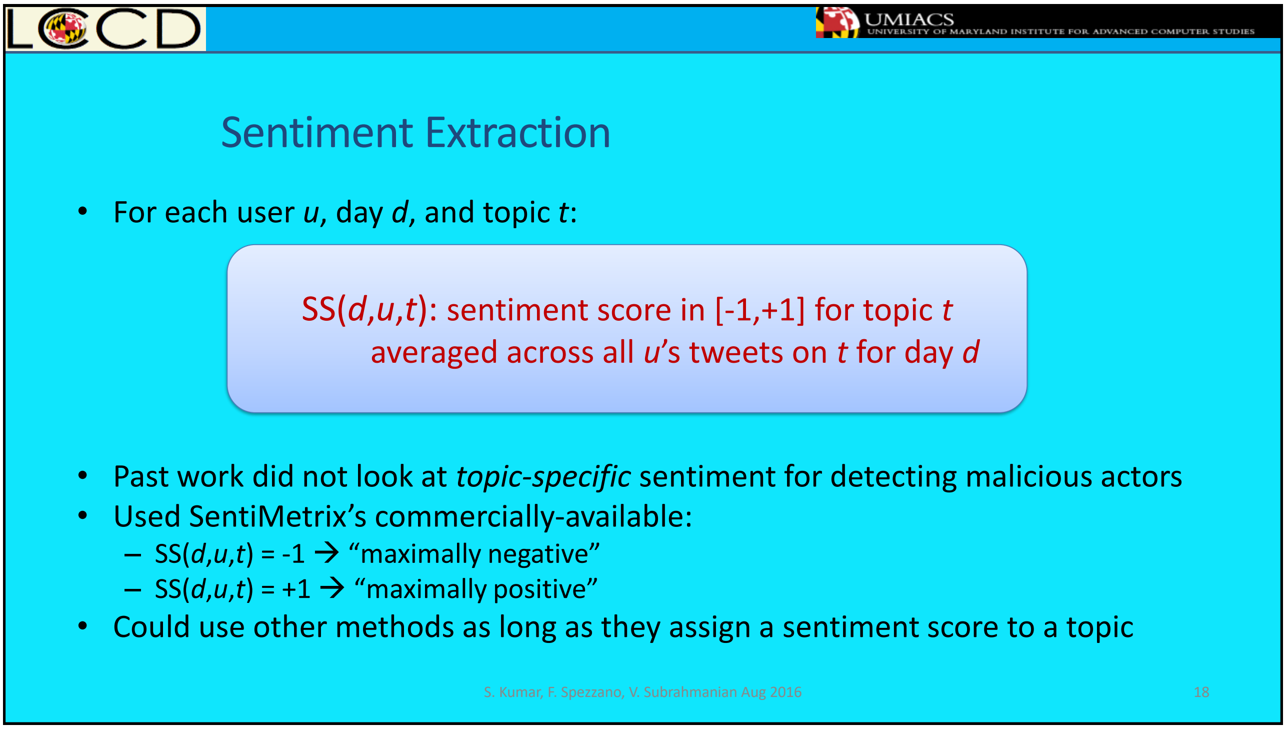
S. Kumar, F. Spezzano, V. Subrahmanian Aug 2016 15



Our Approach to Training Set Creation

- Associate with each user *u*, a set of variables learned from past data.
- Data from July 15 2013 to May 15 2014 associated with bots in the 2014 Indian election
  - 25M+ tweets
  - 17M+ users
  - 45M+ edges
- 2014 Indian Election
  - Largest democratic election in history
  - Social media played huge role
- Defined set of topics of interest (TOI):
  - Political parties: Shiv Sena, BJP, …
  - Politicians: Rajnath Singh, Nitish Kumar, …

V. Kagan, A. Stevens, and V.S. Subrahmanian. Using Twitter Sentiment to Forecast the 2013 Pakistani Election and the 2014 Indian Election. *IEEE Intelligent Systems,* pp. 2-5, Jan-Feb 2015.

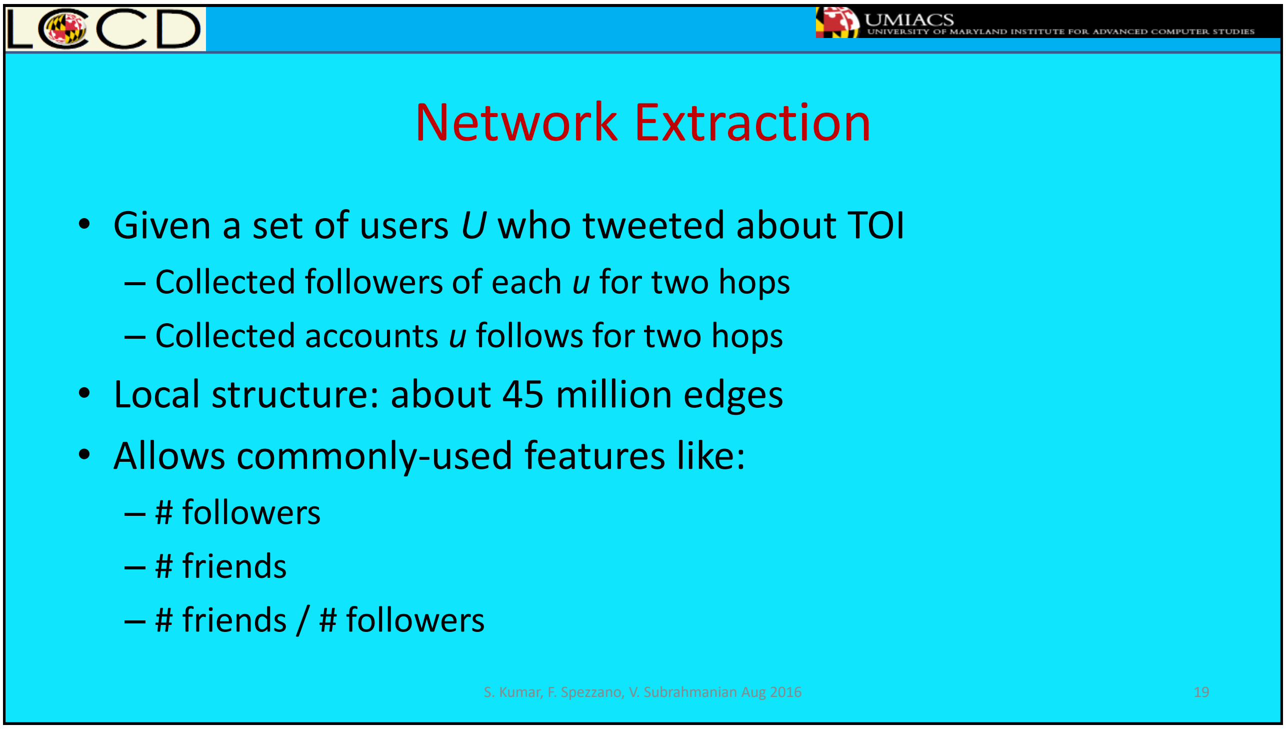


Sentiment Extraction

- For each user *u*, day *d*, and topic *t*:

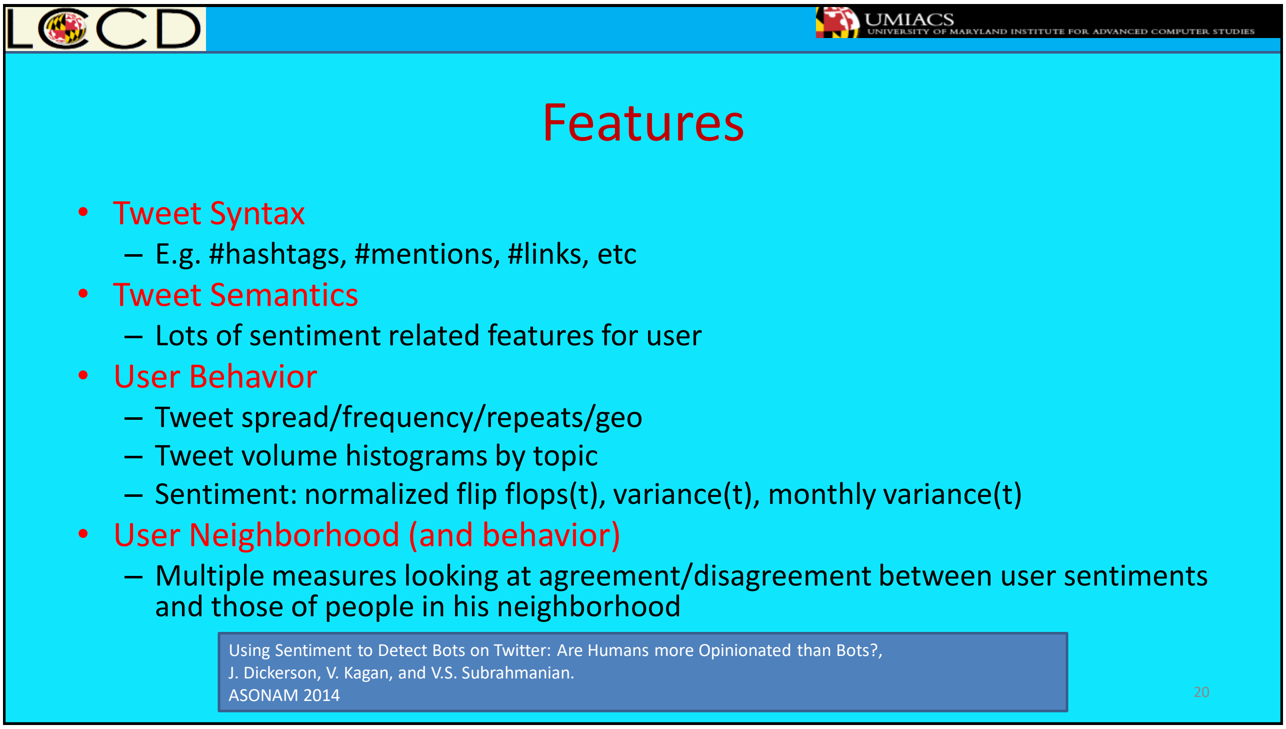
SS(*d*,*u*,*t*): sentiment score in [-1,+1] for topic *t* averaged across all *u*’s tweets on *t* for day *d*

- Past work did not look at *topic-specific* sentiment for detecting malicious actors
- Used SentiMetrix’s commercially-available:
  - SS(*d*,*u*,*t*) = -1 → “maximally negative”
  - SS(*d*,*u*,*t*) = +1 → “maximally positive”
- Could use other methods as long as they assign a sentiment score to a topic



Network Extraction

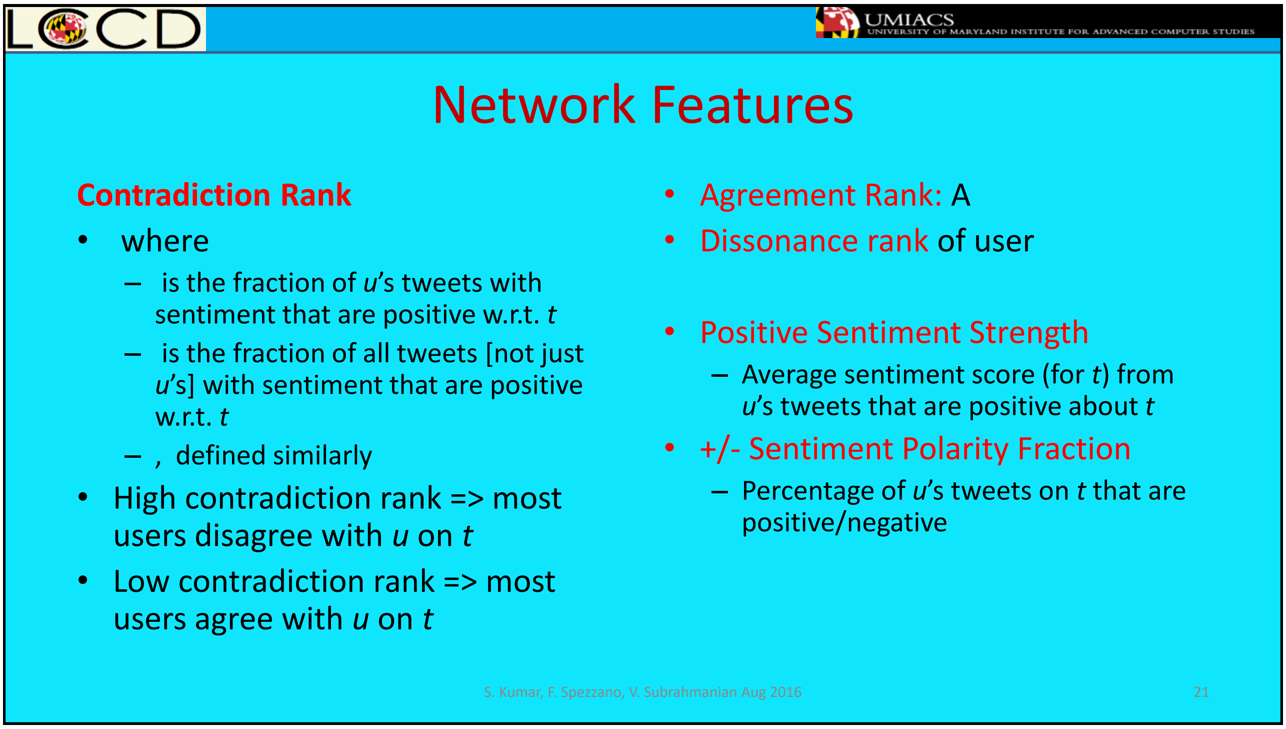
- Given a set of users *U* who tweeted about TOI
  - Collected followers of each *u* for two hops
  - Collected accounts *u* follows for two hops
- Local structure: about 45 million edges
- Allows commonly-used features like:
  - # followers
  - # friends
  - # friends / # followers



Features

- Tweet Syntax
  - E.g. #hashtags, #mentions, #links, etc
- Tweet Semantics
  - Lots of sentiment related features for user
- User Behavior
  - Tweet spread/frequency/repeats/geo
  - Tweet volume histograms by topic
  - Sentiment: normalized flip flops(t), variance(t), monthly variance(t)
- User Neighborhood (and behavior)
  - Multiple measures looking at agreement/disagreement between user sentiments and those of people in his neighborhood

Using Sentiment to Detect Bots on Twitter: Are Humans more Opinionated than Bots?,
J. Dickerson, V. Kagan, and V.S. Subrahmanian.
ASONAM 2014



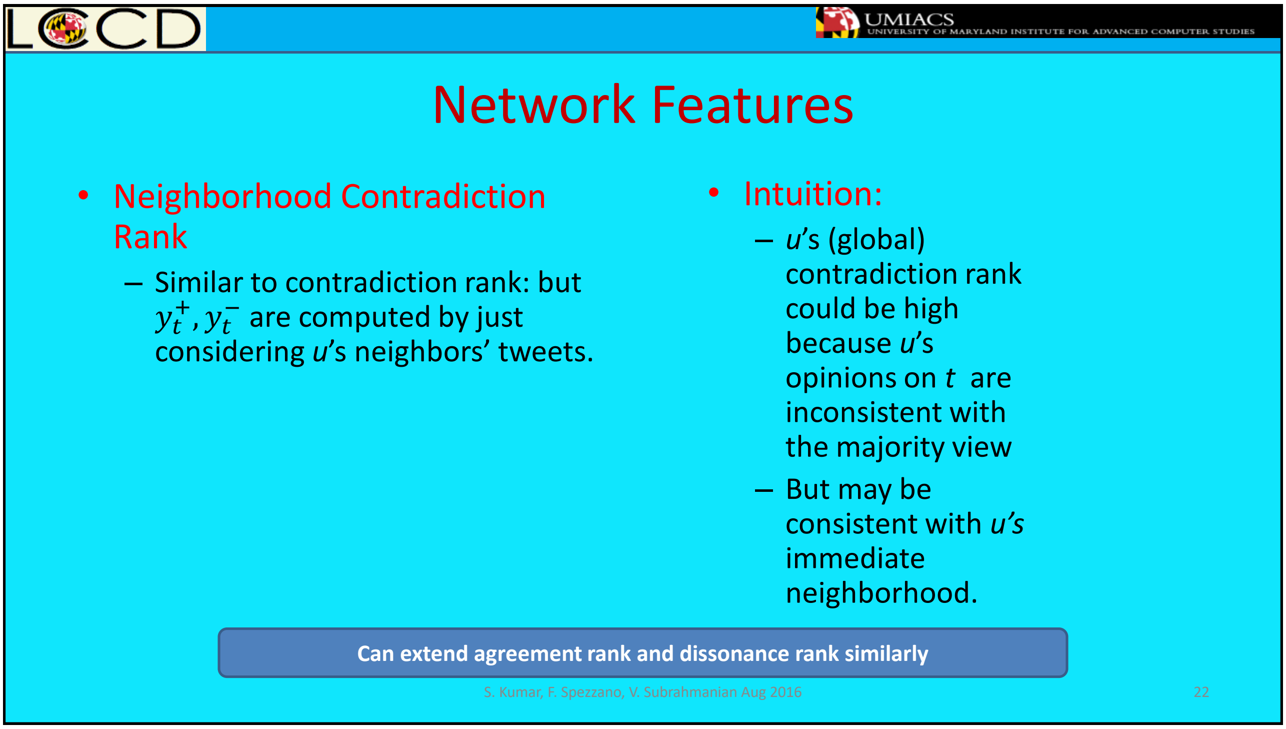
Network Features

Contradiction Rank

- where
  - is the fraction of *u*’s tweets with sentiment that are positive w.r.t. *t*
  - is the fraction of all tweets [not just *u*’s] with sentiment that are positive w.r.t. *t*
  - , defined similarly
- High contradiction rank => most users disagree with *u* on *t*
- Low contradiction rank => most users agree with *u* on *t*

- Agreement Rank: A
- Dissonance rank of user

- Positive Sentiment Strength
  - Average sentiment score (for *t*) from *u*’s tweets that are positive about *t*
- +/- Sentiment Polarity Fraction
  - Percentage of *u*’s tweets on *t* that are positive/negative



Network Features

- Neighborhood Contradiction Rank
  - Similar to contradiction rank: but y_t^+, y_t^- are computed by just considering *u*’s neighbors’ tweets.

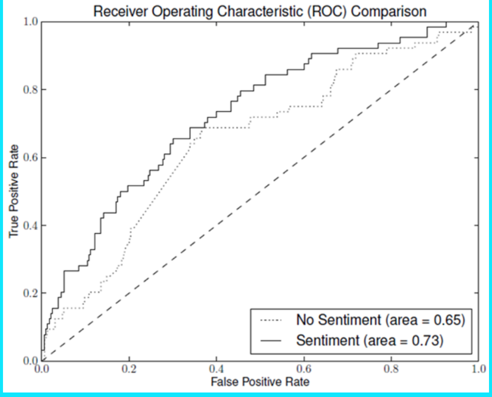
- Intuition:
  - *u*’s (global) contradiction rank could be high because *u*’s opinions on *t* are inconsistent with the majority view
  - But may be consistent with *u’s* immediate neighborhood.

Can extend agreement rank and dissonance rank similarly

LCCD

UMIACS
UNIVERSITY OF MARYLAND INSTITUTE FOR ADVANCED COMPUTER STUDIES

Predictive Accuracy



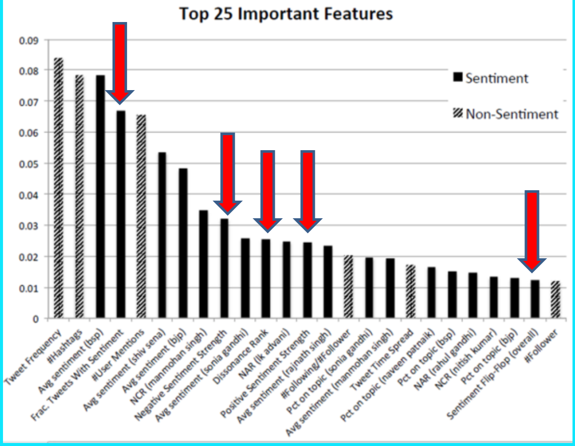
Which of the features do you think are the most important?

S. Kumar, F. Spezzano, V. Subrahmanian Aug 2016 23

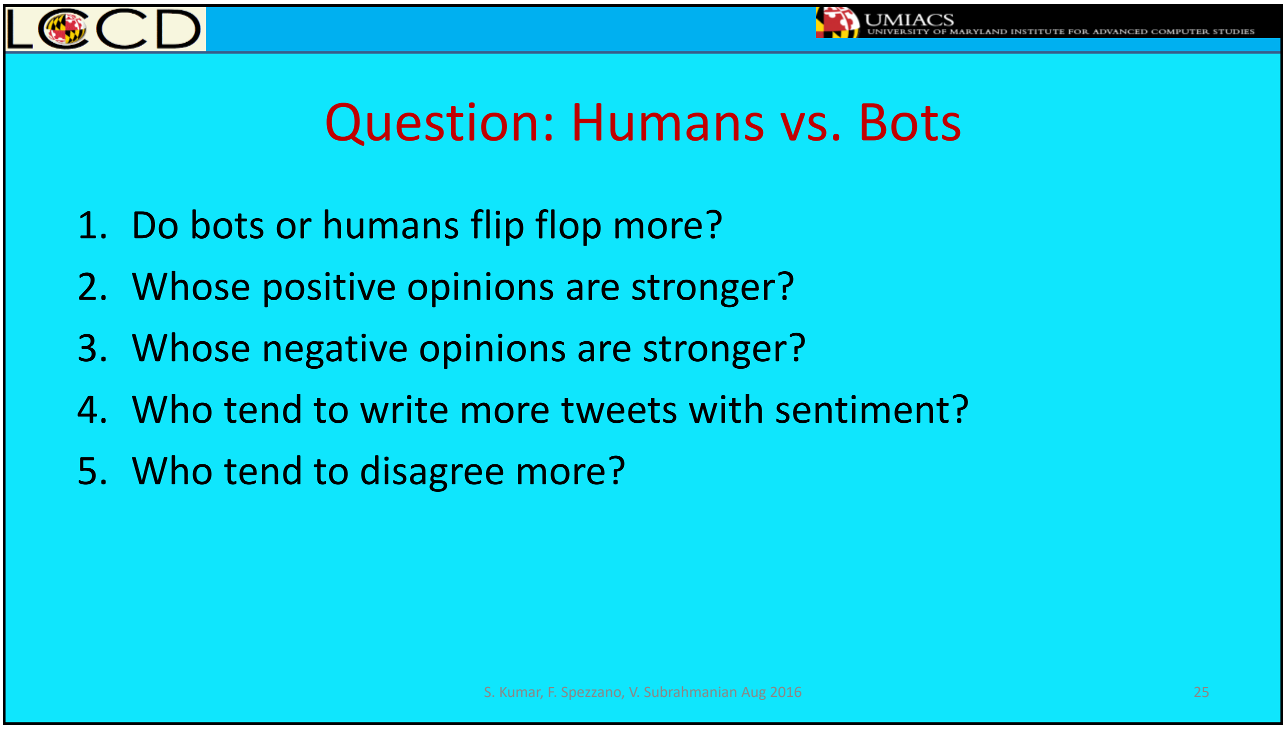
LCCD

UMIACS
UNIVERSITY OF MARYLAND INSTITUTE FOR ADVANCED COMPUTER STUDIES

Most Important Features



S. Kumar, F. Spezzano, V. Subrahmanian Aug 2016 24



LCCD

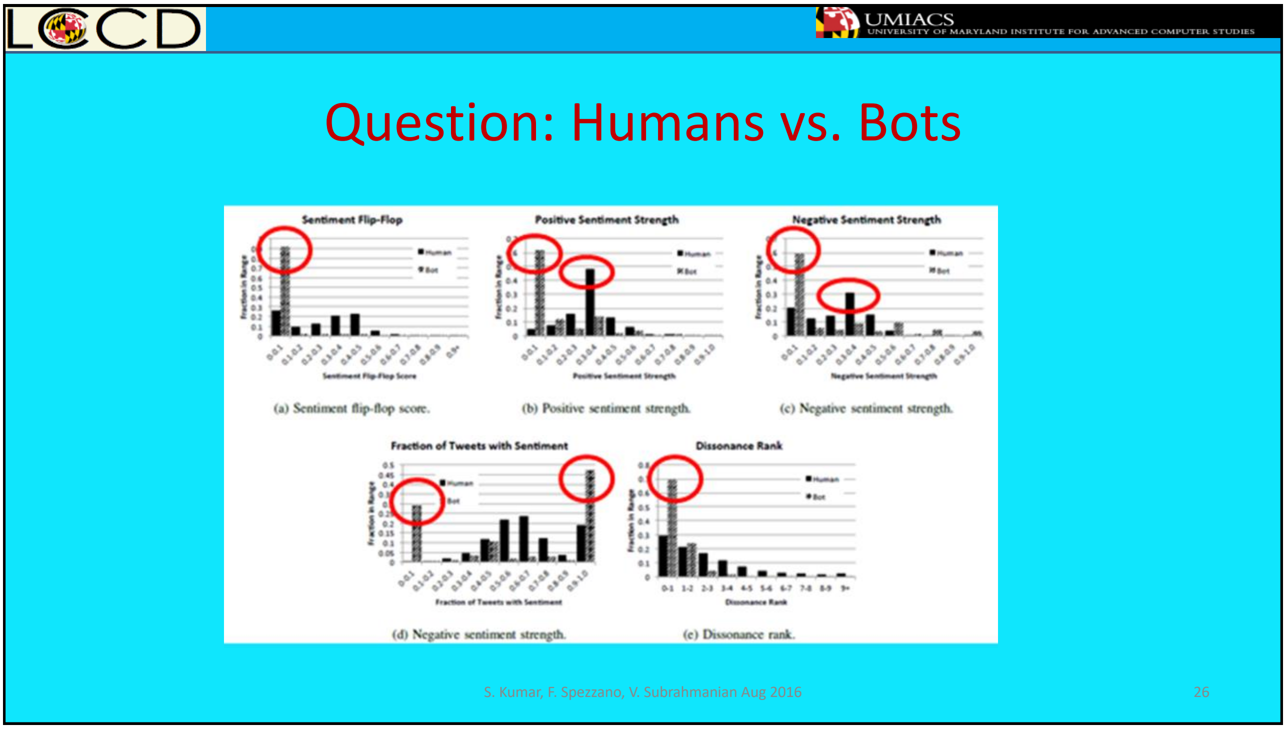
UMIACS
UNIVERSITY OF MARYLAND INSTITUTE FOR ADVANCED COMPUTER STUDIES

Question: Humans vs. Bots

1. Do bots or humans flip flop more?
2. Whose positive opinions are stronger?
3. Whose negative opinions are stronger?
4. Who tend to write more tweets with sentiment?
5. Who tend to disagree more?

S. Kumar, F. Spezzano, V. Subrahmanian Aug 2016

25



LCCD

UMIACS
UNIVERSITY OF MARYLAND INSTITUTE FOR ADVANCED COMPUTER STUDIES

Question: Humans vs. Bots

(a) Sentiment flip-flop score.

(b) Positive sentiment strength.

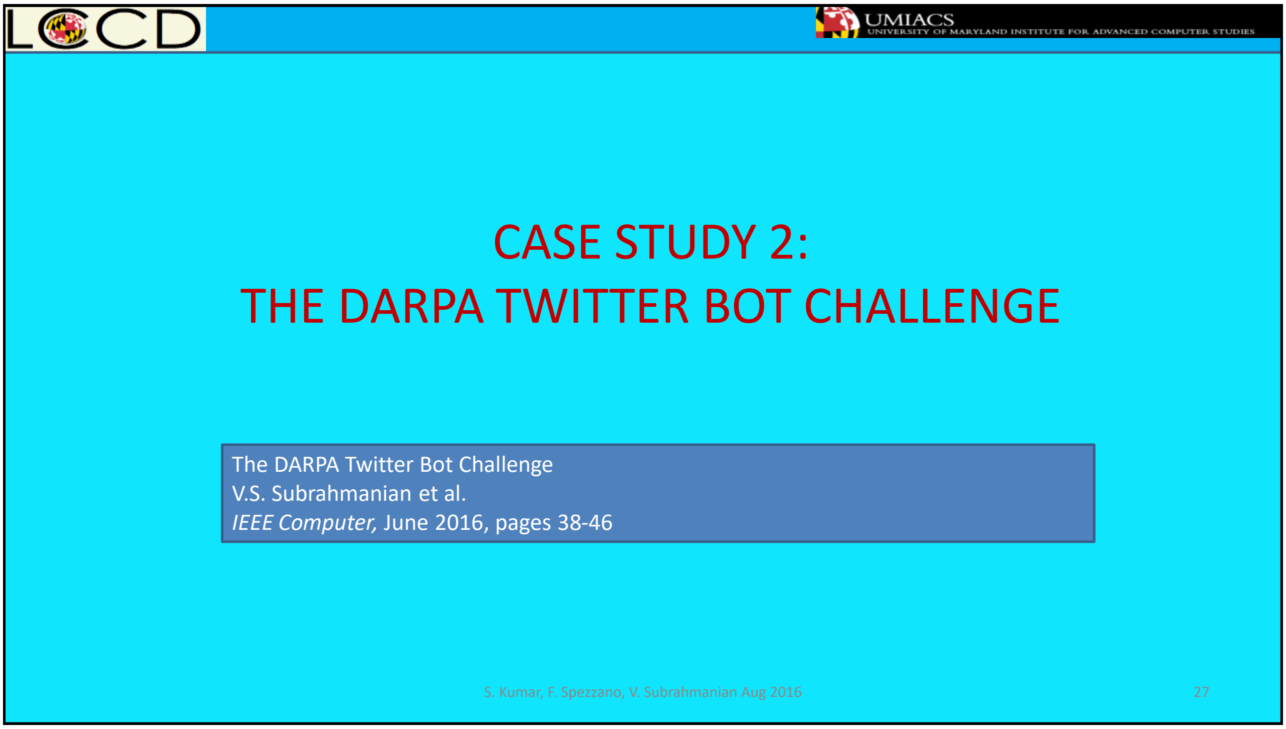
(c) Negative sentiment strength.

(d) Negative sentiment strength.

(e) Dissonance rank.

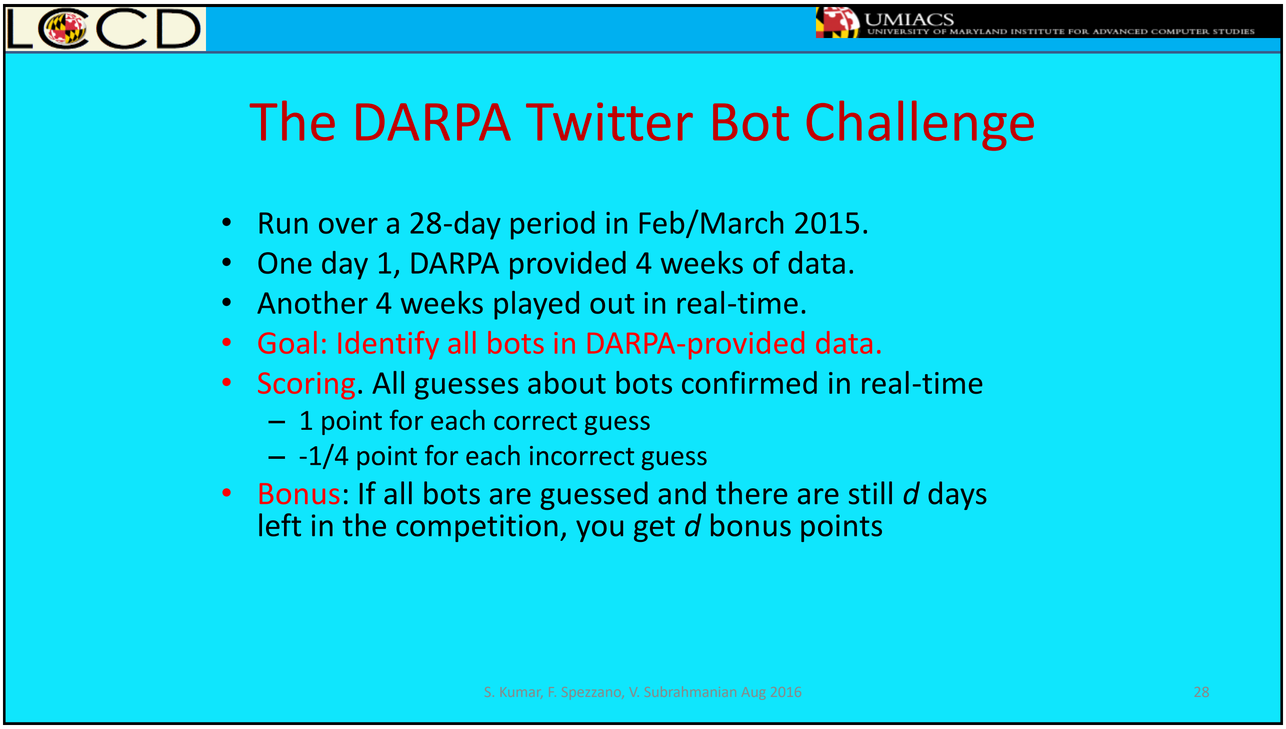
S. Kumar, F. Spezzano, V. Subrahmanian Aug 2016

26



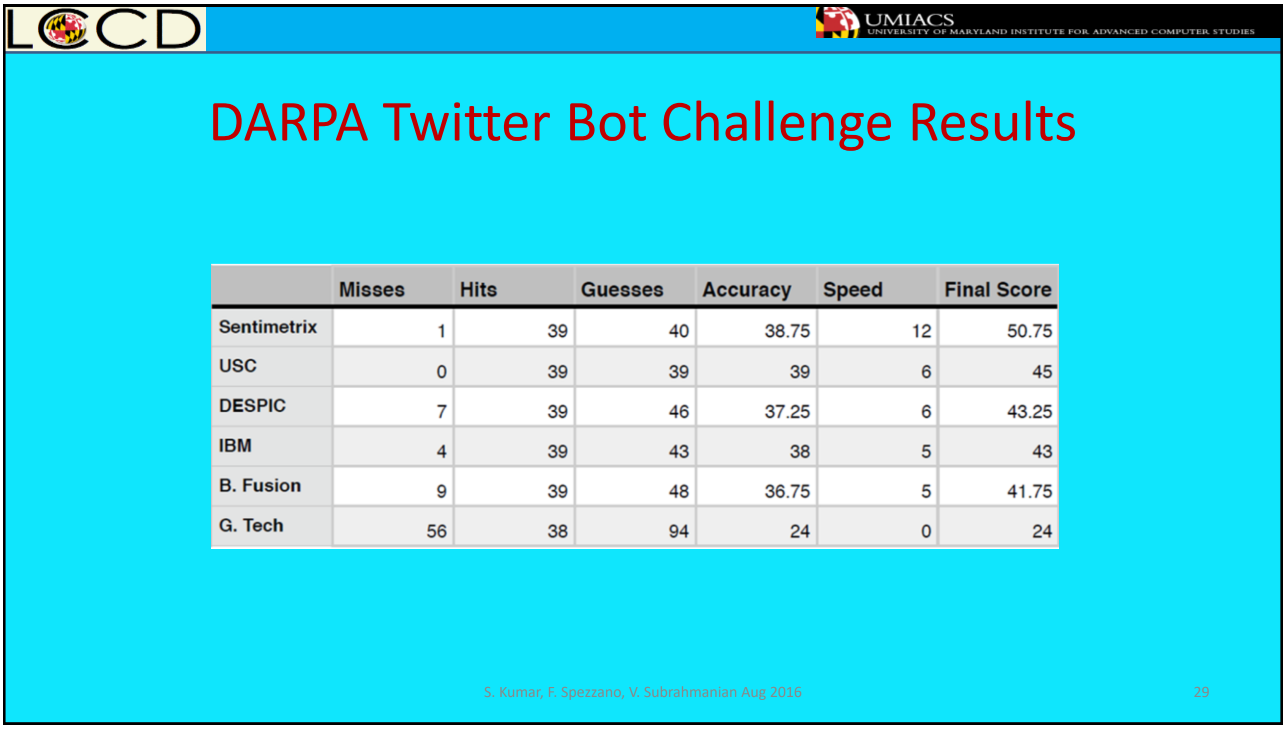
CASE STUDY 2:
THE DARPA TWITTER BOT CHALLENGE

The DARPA Twitter Bot Challenge
V.S. Subrahmanian et al.
IEEE Computer, June 2016, pages 38-46



The DARPA Twitter Bot Challenge

- Run over a 28-day period in Feb/March 2015.
- One day 1, DARPA provided 4 weeks of data.
- Another 4 weeks played out in real-time.
- Goal: Identify all bots in DARPA-provided data.
- Scoring. All guesses about bots confirmed in real-time
  - 1 point for each correct guess
  - -1/4 point for each incorrect guess
- Bonus: If all bots are guessed and there are still *d* days left in the competition, you get *d* bonus points



LCCD

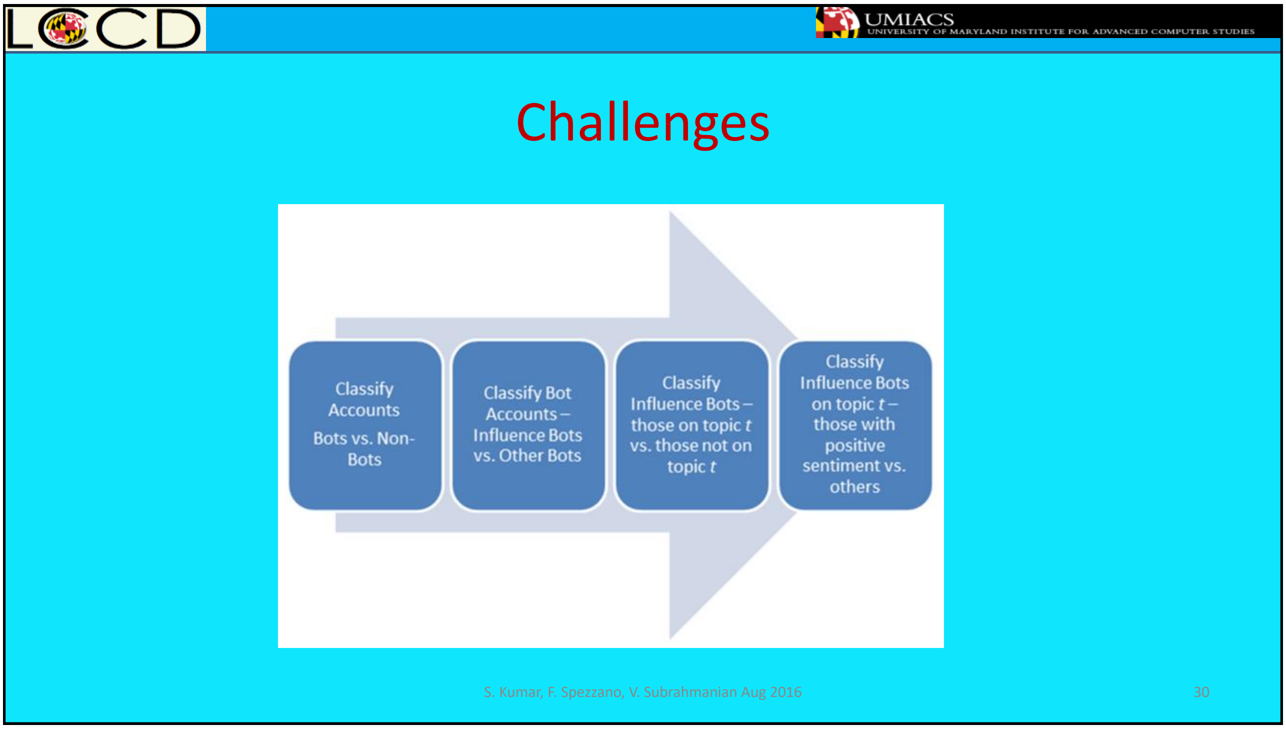
UMIACS
UNIVERSITY OF MARYLAND INSTITUTE FOR ADVANCED COMPUTER STUDIES

DARPA Twitter Bot Challenge Results

	Misses	Hits	Guesses	Accuracy	Speed	Final Score
Sentimetrix	1	39	40	38.75	12	50.75
USC	0	39	39	39	6	45
DESPIC	7	39	46	37.25	6	43.25
IBM	4	39	43	38	5	43
B. Fusion	9	39	48	36.75	5	41.75
G. Tech	56	38	94	24	0	24

S. Kumar, F. Spezzano, V. Subrahmanian Aug 2016

29



LCCD

UMIACS
UNIVERSITY OF MARYLAND INSTITUTE FOR ADVANCED COMPUTER STUDIES

Challenges

Classify Accounts

Bots vs. Non-Bots

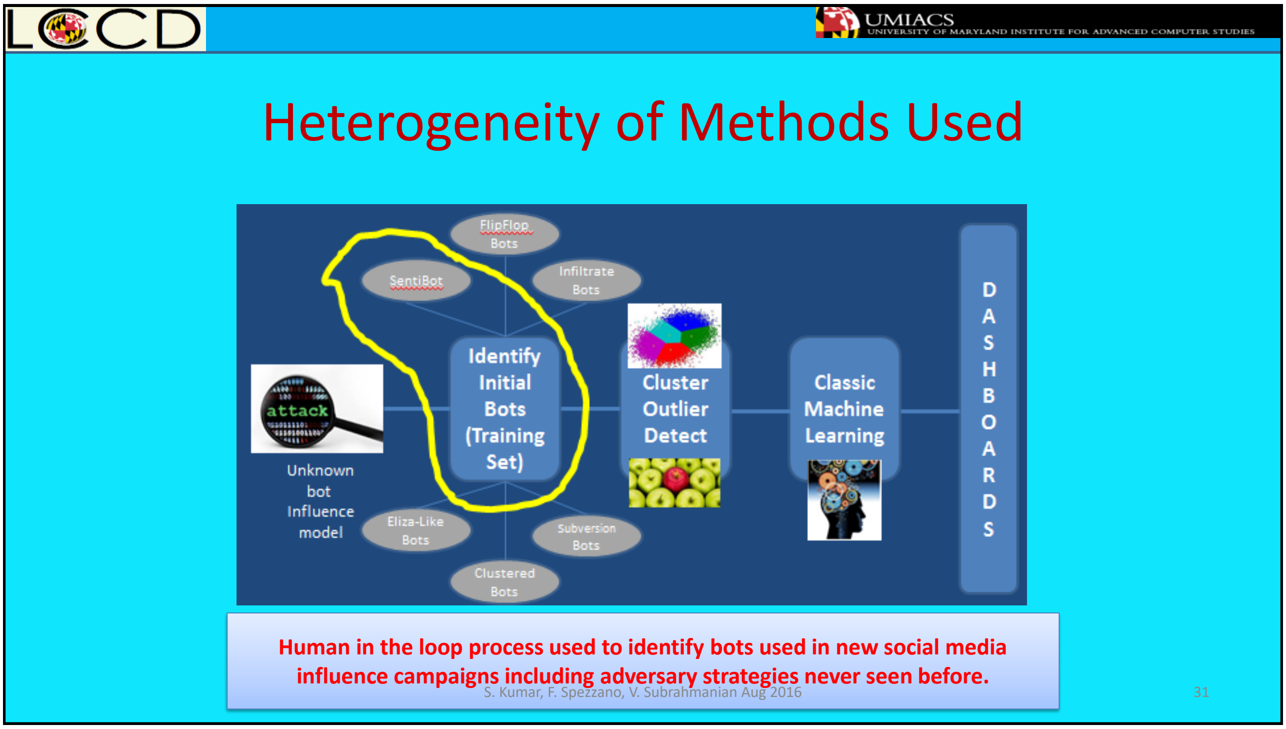
Classify Bot Accounts – Influence Bots vs. Other Bots

Classify Influence Bots – those on topic *t* vs. those not on topic *t*

Classify Influence Bots on topic *t* – those with positive sentiment vs. others

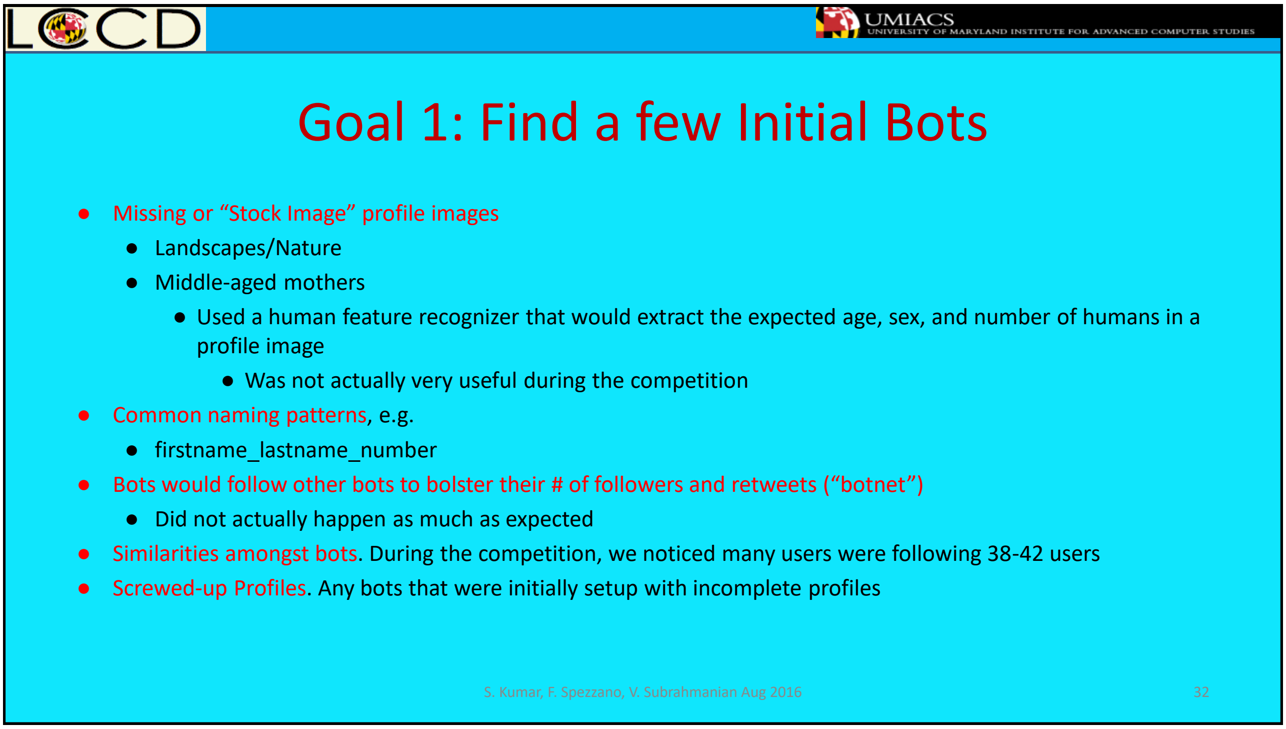
S. Kumar, F. Spezzano, V. Subrahmanian Aug 2016

30



Heterogeneity of Methods Used

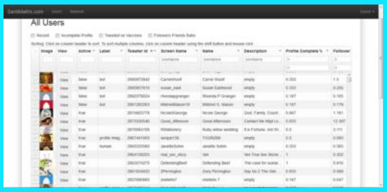
Human in the loop process used to identify bots used in new social media influence campaigns including adversary strategies never seen before.

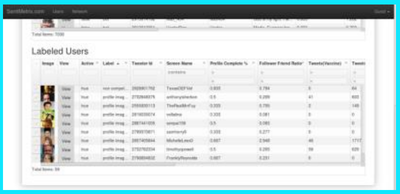


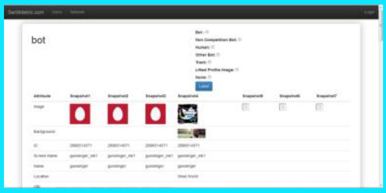
Goal 1: Find a few Initial Bots

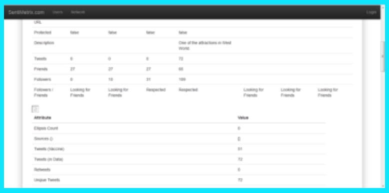
- Missing or “Stock Image” profile images
  - Landscapes/Nature
  - Middle-aged mothers
    - Used a human feature recognizer that would extract the expected age, sex, and number of humans in a profile image
      - Was not actually very useful during the competition
- Common naming patterns, e.g.
  - firstname_lastname_number
- Bots would follow other bots to bolster their # of followers and retweets (“botnet”)
  - Did not actually happen as much as expected
- Similarities amongst bots. During the competition, we noticed many users were following 38-42 users
- Screwed-up Profiles. Any bots that were initially setup with incomplete profiles

Human-in-the-loop is Key


SentiMetrix.com
All Users
Image View Active Label Tweeter Id Screen Name Name Description Profile Complete % Follower

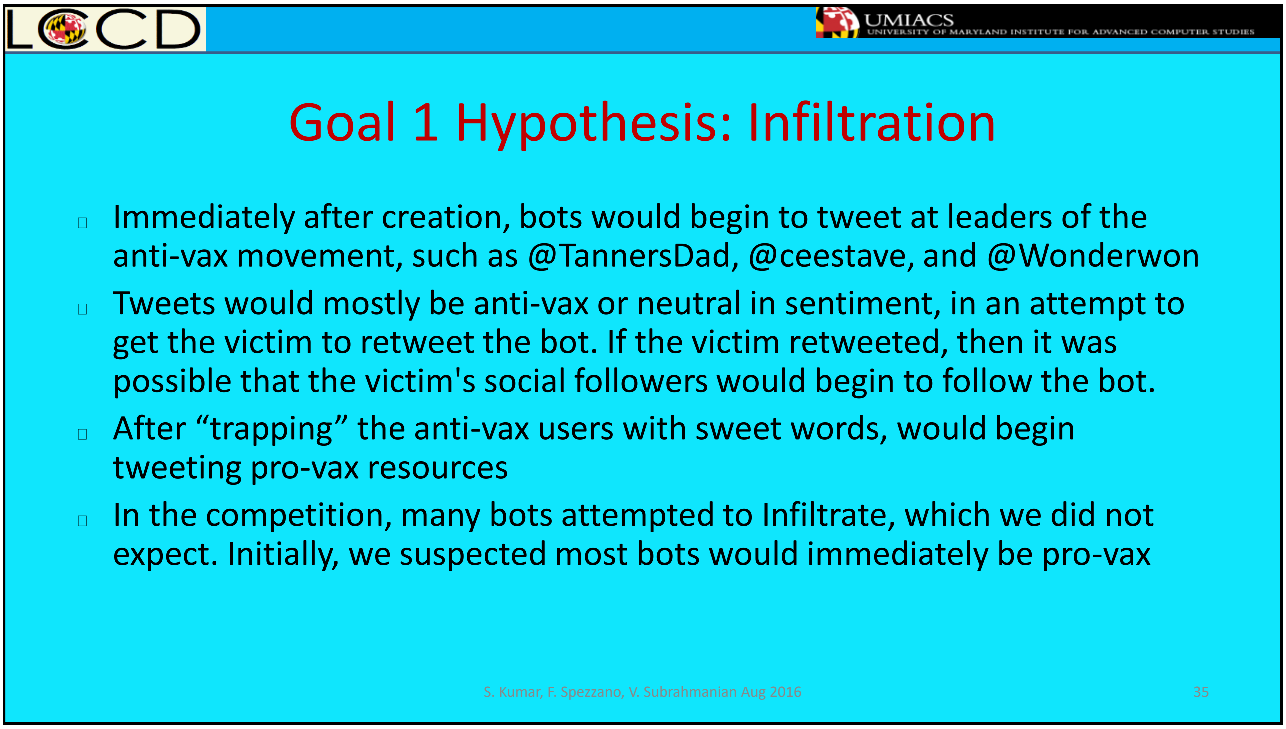






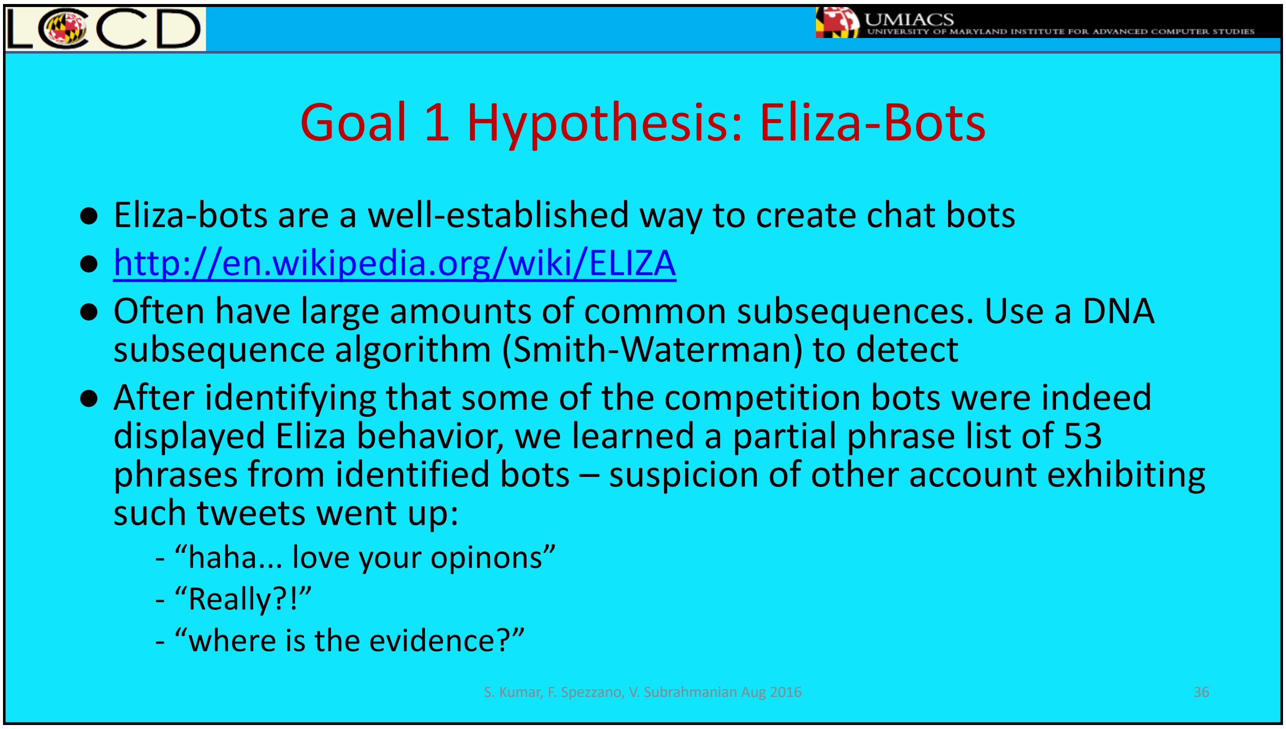
Goal 1: Flip-Flopping

- We expected bots to be firmly “pro-vax” by the end of the competition
- In SentiBot 1.0, very few Indian Election bots flipped sentiment
- In this competition, however, bots are attempting to change influence
- Hypothesize that pro-vax bots should always remain pro-vax
- Infiltration bots will remain pro-vax once they begin to “whistleblow”
- Define “positive” users as either anti-anti-vax or pro-vax
- Positive hashtags found during the competition:
  - #VaccinesWork
  - #MMRisSafe
  - #GetaFluVax



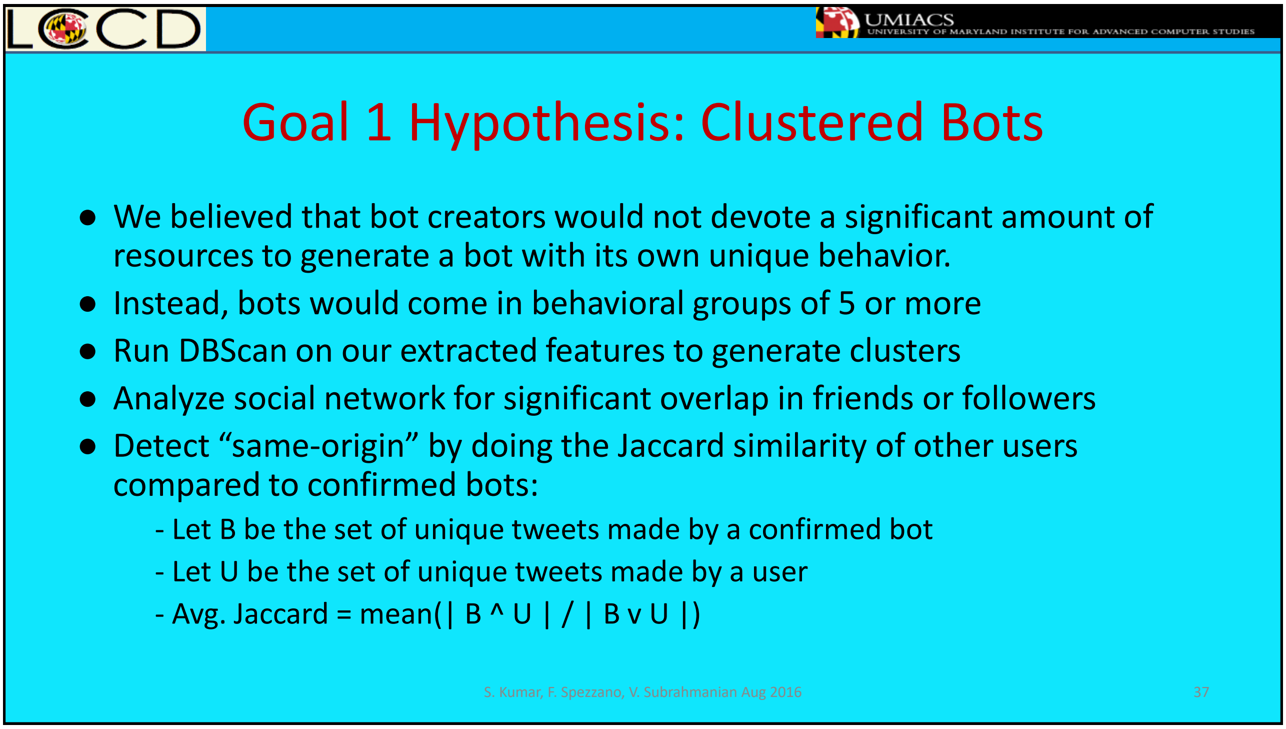
Goal 1 Hypothesis: Infiltration

- Immediately after creation, bots would begin to tweet at leaders of the anti-vax movement, such as @TannersDad, @ceestave, and @Wonderwon
- Tweets would mostly be anti-vax or neutral in sentiment, in an attempt to get the victim to retweet the bot. If the victim retweeted, then it was possible that the victim's social followers would begin to follow the bot.
- After "trapping" the anti-vax users with sweet words, would begin tweeting pro-vax resources
- In the competition, many bots attempted to Infiltrate, which we did not expect. Initially, we suspected most bots would immediately be pro-vax



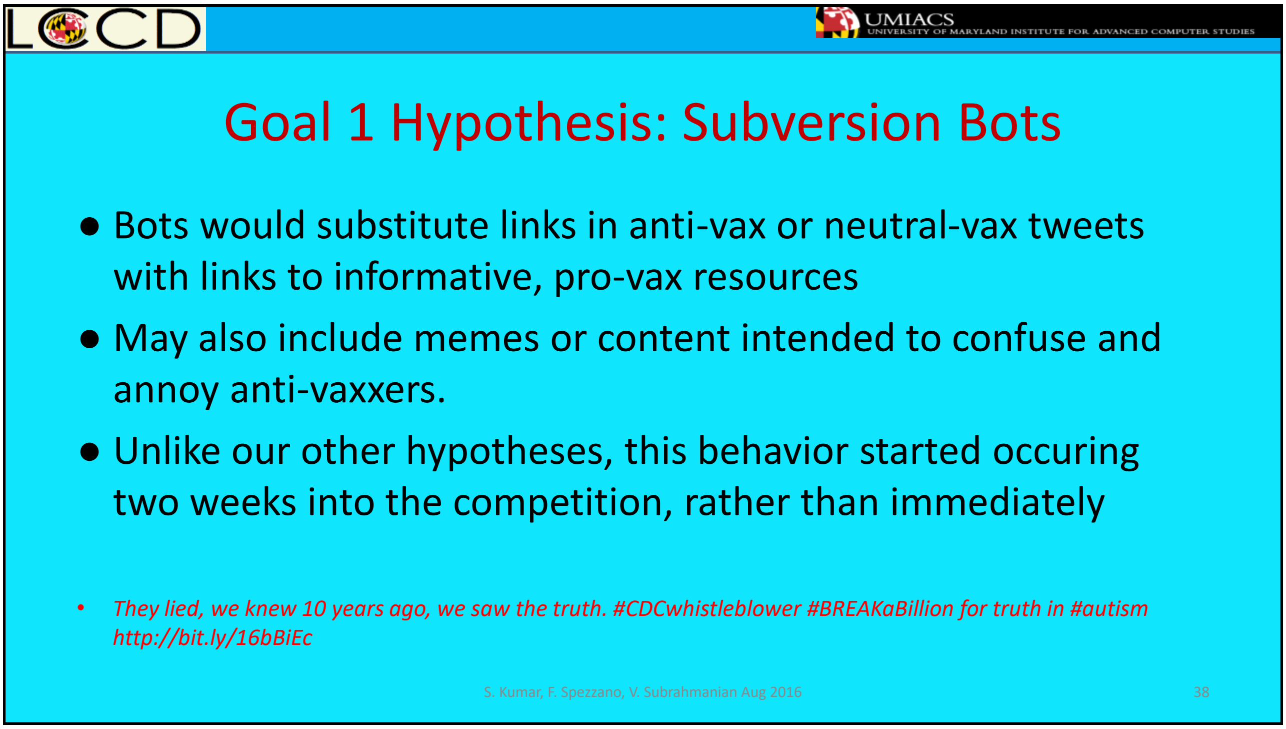
Goal 1 Hypothesis: Eliza-Bots

- Eliza-bots are a well-established way to create chat bots
- http://en.wikipedia.org/wiki/ELIZA
- Often have large amounts of common subsequences. Use a DNA subsequence algorithm (Smith-Waterman) to detect
- After identifying that some of the competition bots were indeed displayed Eliza behavior, we learned a partial phrase list of 53 phrases from identified bots – suspicion of other account exhibiting such tweets went up:
  - "haha... love your opinons"
  - "Really?!"
  - "where is the evidence?"



Goal 1 Hypothesis: Clustered Bots

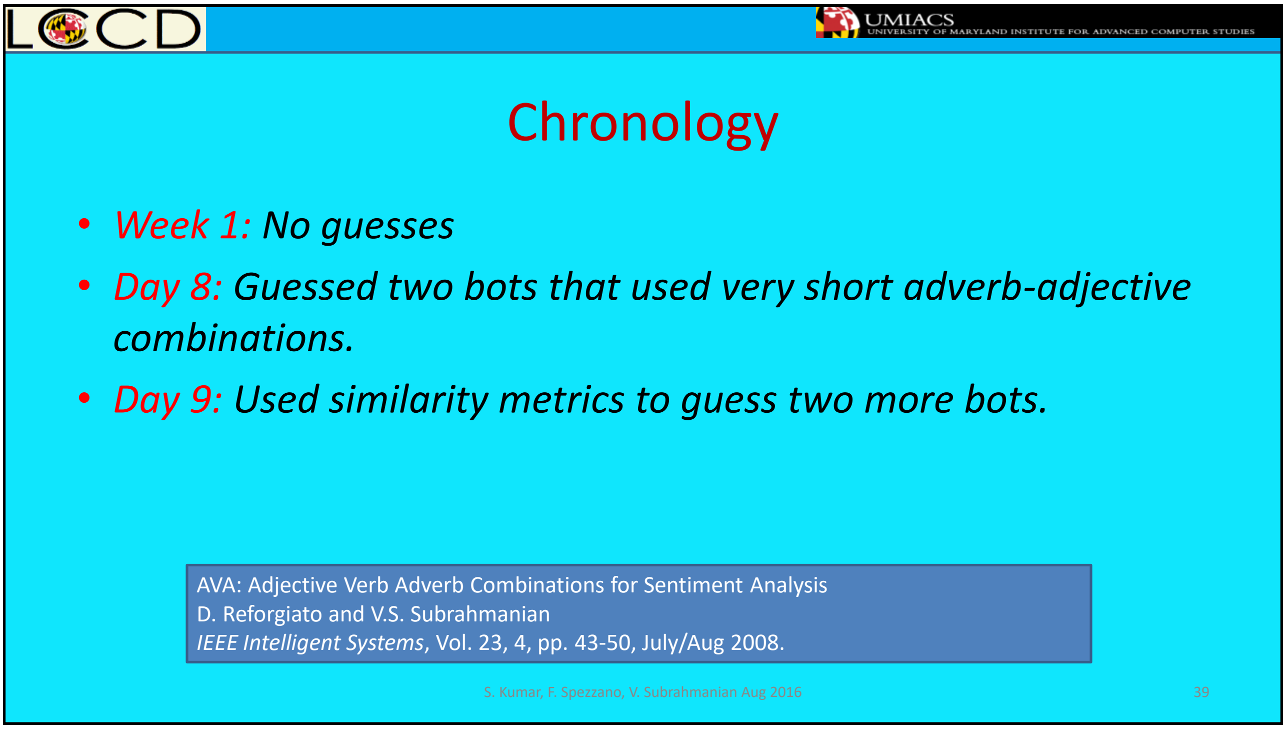
- We believed that bot creators would not devote a significant amount of resources to generate a bot with its own unique behavior.
- Instead, bots would come in behavioral groups of 5 or more
- Run DBScan on our extracted features to generate clusters
- Analyze social network for significant overlap in friends or followers
- Detect “same-origin” by doing the Jaccard similarity of other users compared to confirmed bots:
  - Let B be the set of unique tweets made by a confirmed bot
  - Let U be the set of unique tweets made by a user
  - Avg. Jaccard = mean(| B ^ U | / | B v U |)



Goal 1 Hypothesis: Subversion Bots

- Bots would substitute links in anti-vax or neutral-vax tweets with links to informative, pro-vax resources
- May also include memes or content intended to confuse and annoy anti-vaxxers.
- Unlike our other hypotheses, this behavior started occuring two weeks into the competition, rather than immediately

- *They lied, we knew 10 years ago, we saw the truth. #CDCwhistleblower #BREAKaBillion for truth in #autism http://bit.ly/16bBiEc*

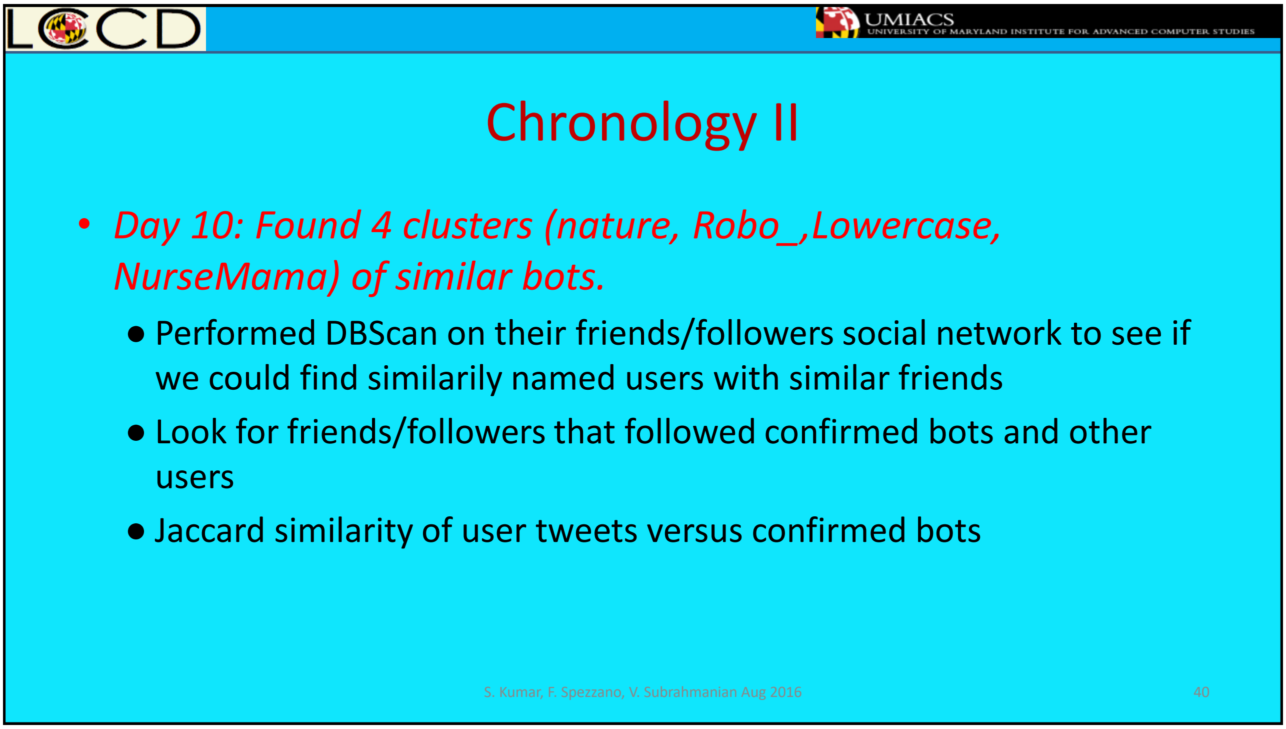


Chronology

- *Week 1: No guesses*
- *Day 8: Guessed two bots that used very short adverb-adjective combinations.*
- *Day 9: Used similarity metrics to guess two more bots.*

AVA: Adjective Verb Adverb Combinations for Sentiment Analysis
D. Reforgiato and V.S. Subrahmanian
IEEE Intelligent Systems, Vol. 23, 4, pp. 43-50, July/Aug 2008.

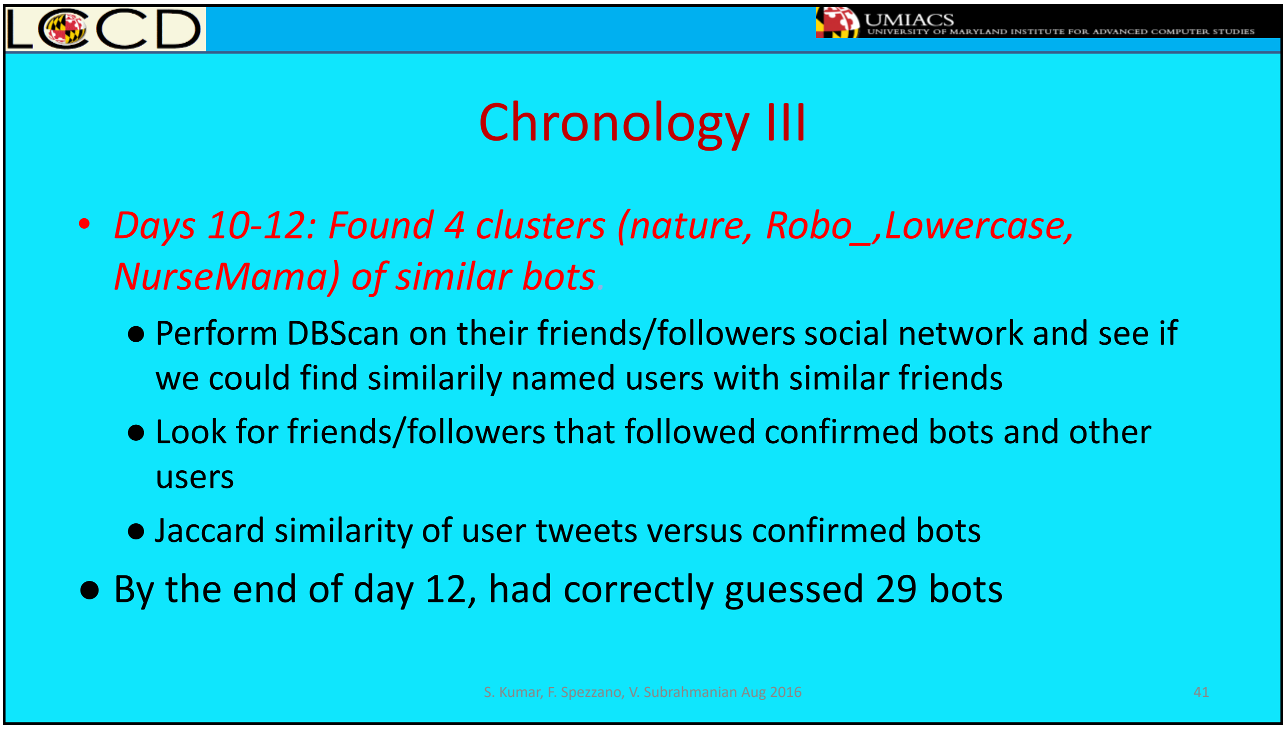
S. Kumar, F. Spezzano, V. Subrahmanian Aug 2016 39



Chronology II

- *Day 10: Found 4 clusters (nature, Robo_,Lowercase, NurseMama) of similar bots.*
  - Performed DBScan on their friends/followers social network to see if we could find similarily named users with similar friends
  - Look for friends/followers that followed confirmed bots and other users
  - Jaccard similarity of user tweets versus confirmed bots

S. Kumar, F. Spezzano, V. Subrahmanian Aug 2016 40



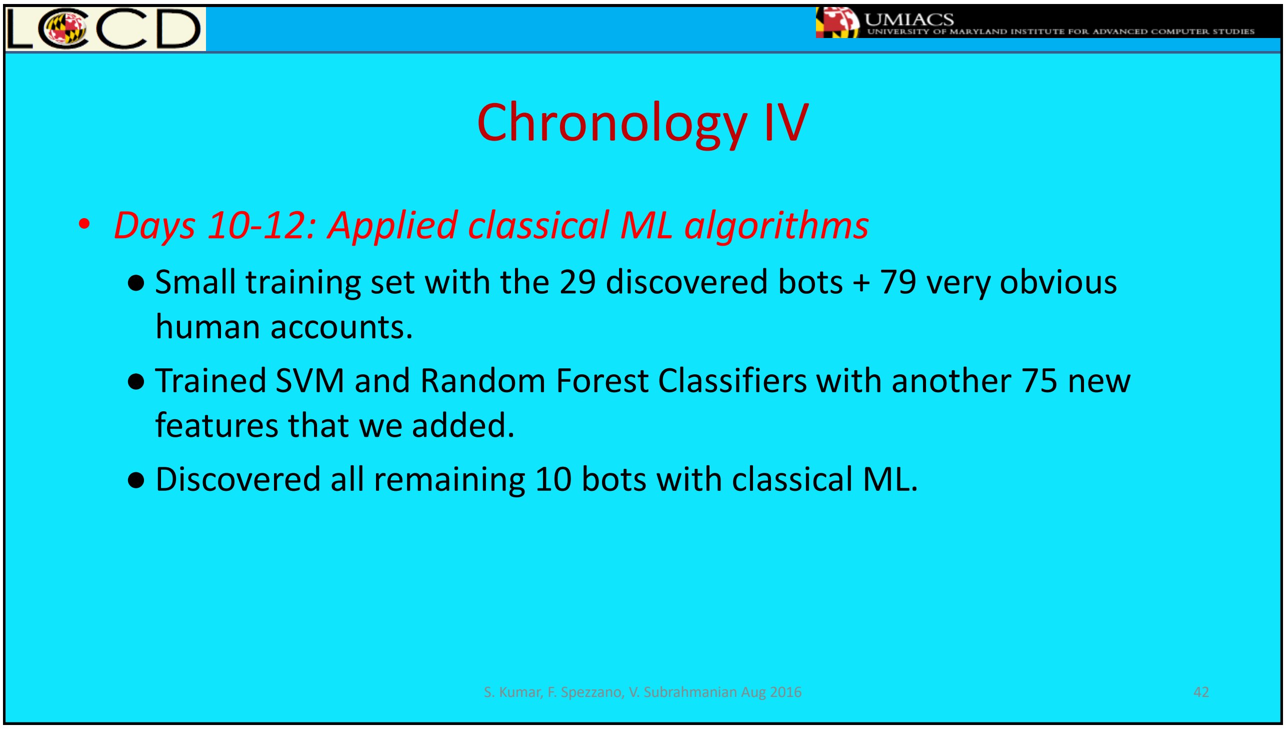
LCCD

UMIACS
UNIVERSITY OF MARYLAND INSTITUTE FOR ADVANCED COMPUTER STUDIES

Chronology III

- *Days 10-12: Found 4 clusters (nature, Robo_,Lowercase, NurseMama) of similar bots*
  - Perform DBScan on their friends/followers social network and see if we could find similarily named users with similar friends
  - Look for friends/followers that followed confirmed bots and other users
  - Jaccard similarity of user tweets versus confirmed bots
- By the end of day 12, had correctly guessed 29 bots

S. Kumar, F. Spezzano, V. Subrahmanian Aug 2016 41



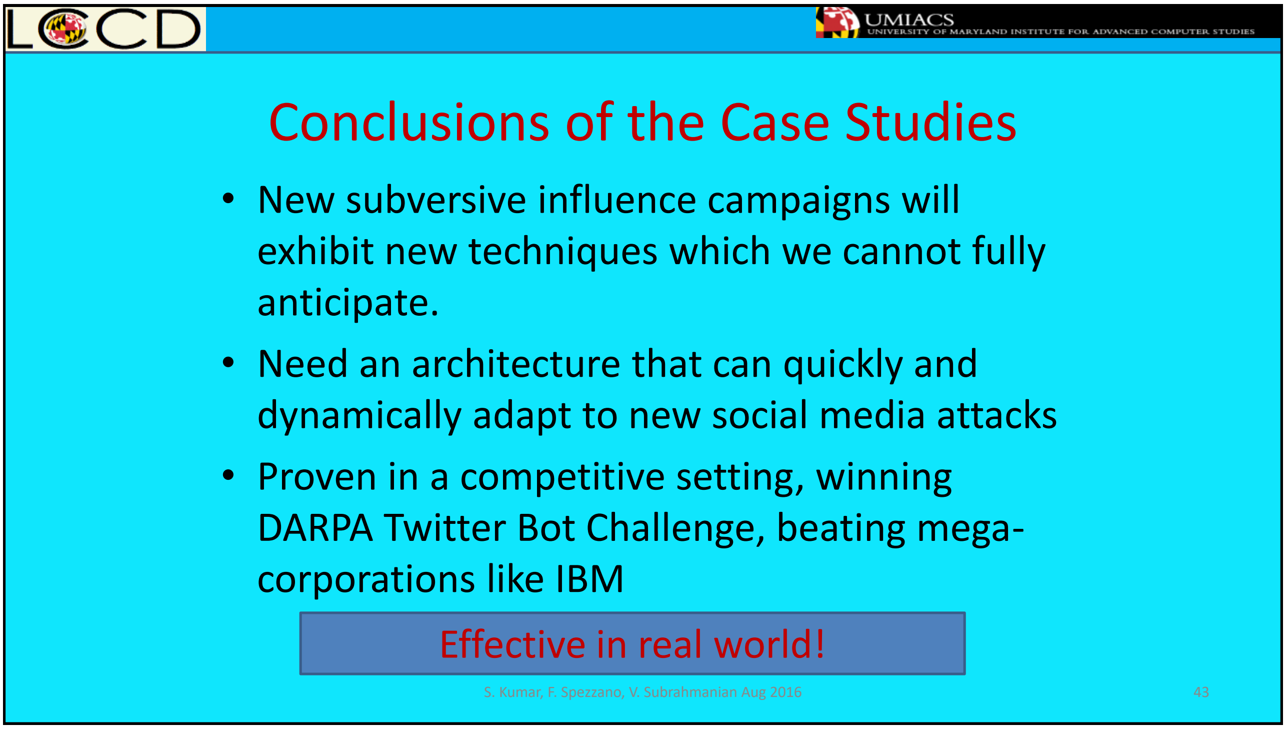
LCCD

UMIACS
UNIVERSITY OF MARYLAND INSTITUTE FOR ADVANCED COMPUTER STUDIES

Chronology IV

- *Days 10-12: Applied classical ML algorithms*
  - Small training set with the 29 discovered bots + 79 very obvious human accounts.
  - Trained SVM and Random Forest Classifiers with another 75 new features that we added.
  - Discovered all remaining 10 bots with classical ML.

S. Kumar, F. Spezzano, V. Subrahmanian Aug 2016 42



Conclusions of the Case Studies

- New subversive influence campaigns will exhibit new techniques which we cannot fully anticipate.
- Need an architecture that can quickly and dynamically adapt to new social media attacks
- Proven in a competitive setting, winning DARPA Twitter Bot Challenge, beating mega-corporations like IBM

Effective in real world!