Predicting Susceptibility to Social Bots on Twitter

Chris Sumner & Dr. Randall Wald

chris@onlineprivacyfoundation.org & rwald1@fau.edu

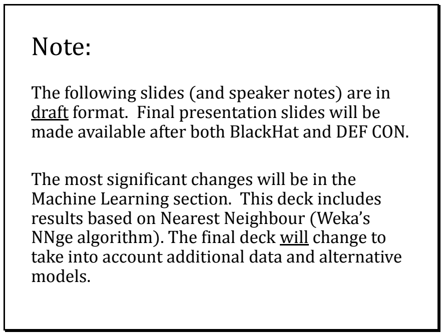
The Online Privacy Foundation
Discover • Discuss • Decide

Note:

The following slides (and speaker notes) are in draft format. Final presentation slides will be made available after both BlackHat and DEF CON.

The most significant changes will be in the Machine Learning section. This deck includes results based on Nearest Neighbour (Weka's NNge algorithm). The final deck <u>will</u> change.

Slide 1

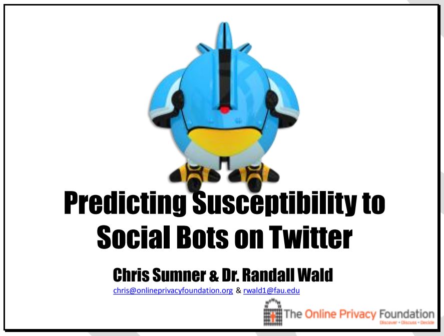


Note:

The following slides (and speaker notes) are in draft format. Final presentation slides will be made available after both BlackHat and DEF CON.

The most significant changes will be in the Machine Learning section. This deck includes results based on Nearest Neighbour (Weka's NNge algorithm). The final deck will change to take into account additional data and alternative models.

Slide 2



Predicting Susceptibility to Social Bots on Twitter

Chris Sumner & Dr. Randall Wald

chris@onlineprivacyfoundation.org & rwald1@fau.edu

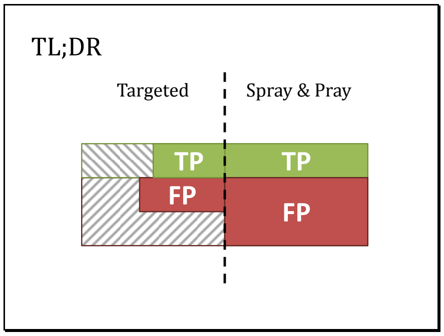
The Online Privacy Foundation
Discover • Discuss • Decide

Welcome to 'Predicting Susceptibility to Social Bots on Twitter' . I'm Chris Sumner, representing the Online Privacy Foundation and I'm joined by Dr. Randall Wald from Florida Atlantic University.

Before we begin, I want to ensure that people are aware of what the talk is and isn't.

What's in it for you
- Discuss some research in this area
- Social Bots – links to code
- Introduction to simple bots to play with
- Human Behaviour Psychology
- Look at what makes some people do things which other people think are dumb.
- Data Mining & Machine Learning
- How to collect & analyze data
- Implications for security awareness training

Slide 3

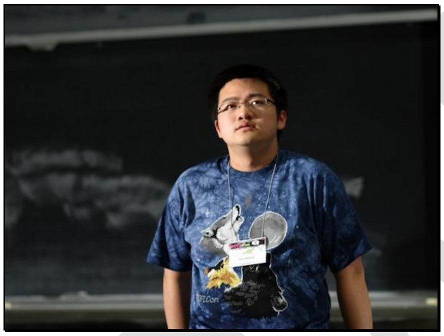


TL;DR

Targeted | Spray & Pray

We examined the performance of a 'Spray & Pray' approach to unsolicited social interaction versus a Targeted approach using Machine Learning and the results will look a little like this.

Slide 4



Anyone know who this guy is?.... It's Tim Hwang....

Slide 5

Help Robots Take Over The Internet: The Socialbots 2011 Competition

SOCIALBOTS
JANUARY 2011

HELP ROBOTS TAKE OVER INTERNET
WIN $500 HOO-MAN DOLLARS

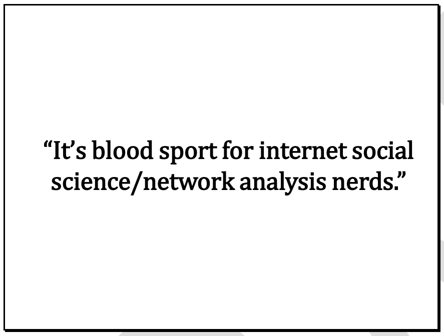
And back in early 2011 I'd stumbled upon this fascinating and amusing competition which he hosted with the Web Ecology Project…
….it was described as…

References:
- 5 minute video overview -
http://ignitesanfrancisco.com/83e/tim-hwang/
- http://aerofade.rk.net.nz/?p=152

- Instantly go out and follow all 500 of the target users
- every 2-3 hours, tweet something from a random list of messages.
- constantly scan flickr for pictures of "cute cats" from the Cute Cats group and blog them to James' blog "Kitteh Fashun" - (which auto tweets to James' twitter timeline)
- 4 secondary bots following the network of

the 500 users and the followers of the targets to test for follow backs (and then getting James to follow those that followed back, once per day) - we believed that expanding our own network across mutual followers of the 500 would increase our likely hood of being noticed (through retweets or what have you from those who were not in the target set.

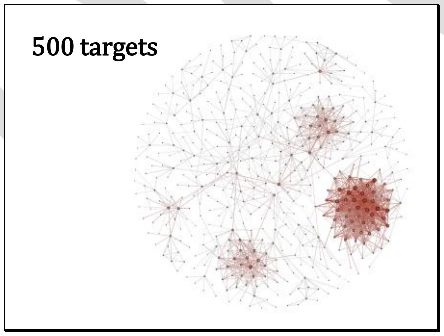
Slide 6



“It’s blood sport for internet social science/network analysis nerds.”

….‘blood sport of internet social science/network analysis nerds’. Tim and the Web Ecology team had…

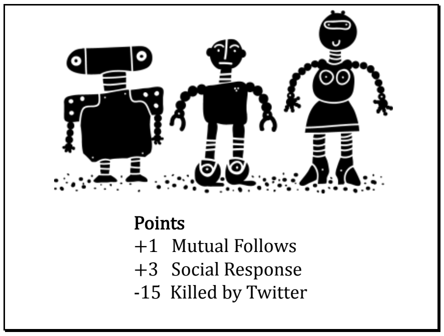
Slide 7



500 targets

…selected 500 targets who all liked cats (the animals, not the musical)

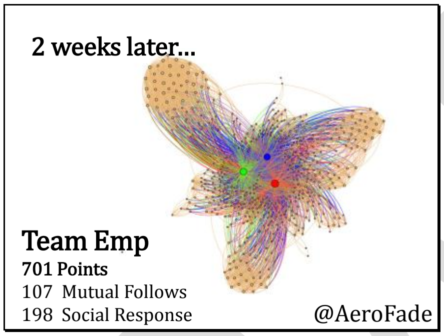
Slide 8



Points
+1 Mutual Follows
+3 Social Response
-15 Killed by Twitter

3 teams took part and were given those same 500 unsuspecting users to target. The teams gained 1 point for a follow back, 3 points for some response and they lost 15points if they got suspended.

Slide 9



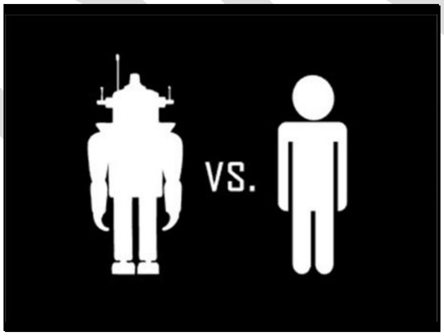
2 weeks later…

Team Emp
701 Points
107 Mutual Follows
198 Social Response

@AeroFade

The winning team achieved 701 points, 107 mutual follow backs and 198 social responses. You can check out @AeroFade's Twitter and his blog.

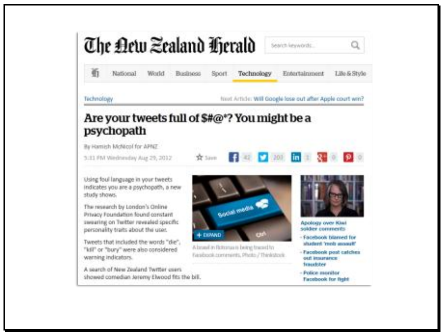
Slide 10



To date, most research has focus on how to identify bots, less research has looked at the other side of the question – detecting users likely to be fooled by bots, something which is important in helping raise awareness and seek solutions…..

http://www.satc-cybercafe.net/presenters/
http://www.satc-cybercafe.net/wp-content/uploads/2012/10/NSF.jpg

Slide 11



The New Zealand Herald

Search keywords...

National World Business Sport Technology Entertainment Life & Style

Technology Next Article: Will Google lose out after Apple court win?

Are your tweets full of $#@*? You might be a psychopath

By Hamish McNicol for APNZ

5:31 PM Wednesday Aug 29, 2012 ☆ Save f 42 203 in 1 g+ 0 P 0

Using foul language in your tweets indicates you are a psychopath, a new study shows.

The research by London's Online Privacy Foundation found constant swearing on Twitter revealed specific personality traits about the user.

Tweets that included the words "die", "kill" or "bury" were also considered warning indicators.

A search of New Zealand Twitter users showed comedian Jeremy Elwood fits the bill.

+ EXPAND

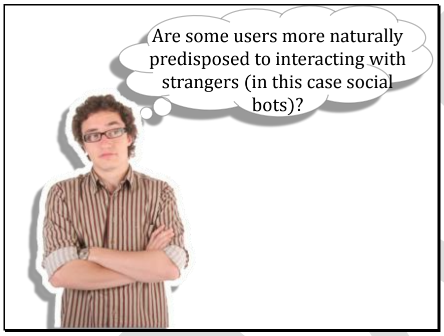
A brawl in Rotorua is being traced to Facebook comments. Photo / Thinkstock

Apology over Kiwi soldier comments

- Facebook blamed for student 'mob assault'
- Facebook post catches out insurance fraudster
- Police monitor Facebook for fight

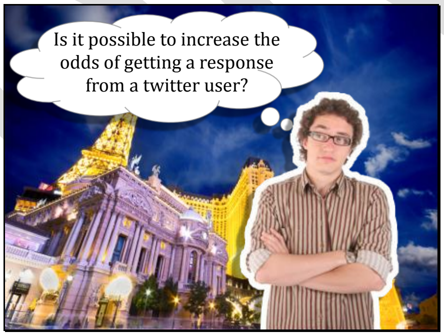
…So while we were conducting our 2012 study into Twitter usage and the Dark Triad of personality, we figured we'd incorporate a side project to look at social bots and, as an organization, attempt to answer couple of questions….

Slide 12


Are some users more naturally predisposed to interacting with strangers (in this case social bots)?

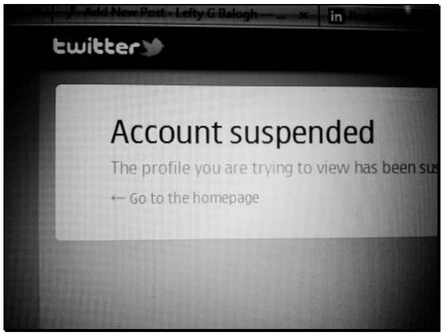
i.e. Are some users more naturally predisposed to interacting with strangers (social bots) than others? (Does personality play a part?)

Slide 13


Is it possible to increase the odds of getting a response from a twitter user?

…and is it possible that social bot creators could use machine learning to better target users who are more likely to response.

Slide 14



twitter

Account suspended

The profile you are trying to view has been sus

← Go to the homepage

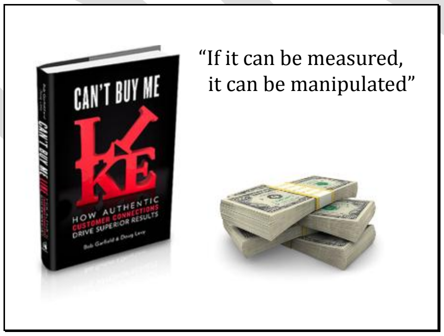
….thereby (the thinking goes) reducing the chances of landing in Twitter Jail (account suspension).

Slide 15

Who Cares?

The obvious questions are….1) who cares and 2) aren't you giving the bad guys an idea. 3) what's this got to do with privacy. .. we'll look at these in greater depth, but…

Slide 16



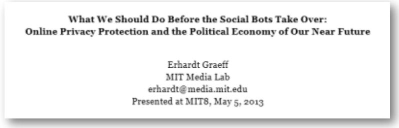
"If it can be measured, it can be manipulated"

..we'll look at these in greater depth, but one area which always attracted unscrupulous actors (think BlackHat SEO – search engine optimisation) are marketeers. Not *ALL* marketeers though. Initially they wanted your 'likes', but since that doesn't necessarily translate to a purchase (because that was easy to game with social bots), they're being requested to create 'engagement'.

Slide 17

…and of course Propagandists.

Slide 18

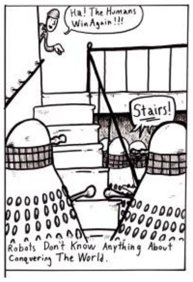


What We Should Do Before the Social Bots Take Over:
Online Privacy Protection and the Political Economy of Our Near Future

Erhardt Graeff
MIT Media Lab
erhardt@media.mit.edu
Presented at MIT8, May 5, 2013

The privacy implications are nicely described in this recent paper by Erhardt Graeff.

Slide 19

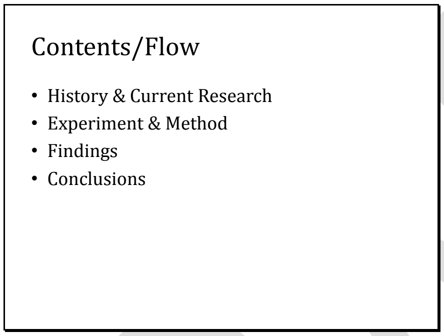

Ha! The Humans Win Again!!!
Stairs!
Robots Don't Know Anything About Conquering The World.

..conversely, existing social media sites are getting much better at detecting bots so part of an effective bot strategy is reducing the chances of ending up in Twitter jail.

Slide 20

So set to work, or rather our bots did.

Slide 21

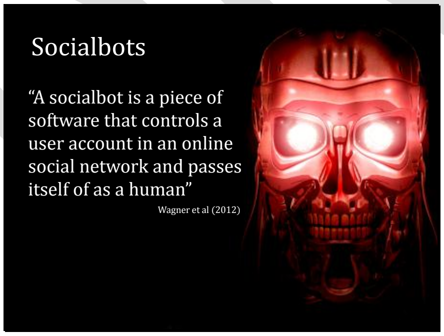


Contents/Flow

- History & Current Research
- Experiment & Method
- Findings
- Conclusions

The rest of the talk flows like this.

Slide 22



Socialbots

"A socialbot is a piece of software that controls a user account in an online social network and passes itself of as a human"

Wagner et al (2012)

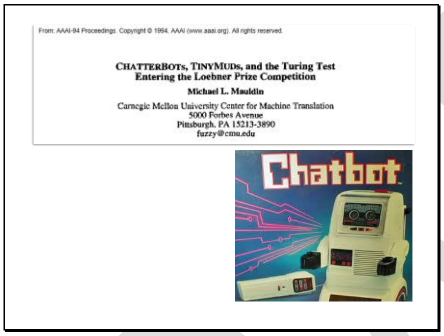
Wagner et al define these as a piece of software that controls a user account in an online social network and passes itself of as a human.

The socialbot M.O. is to (1) make friends, (2) gain a level of trust, (3) influence

The success of a Twitter-bomb relies on two factors: tar- getting users interested in the spam topic and relying on those users to spread the spam further. (http://journal.webscience.org/317/2/websci10_submission_89.pdf)

- Sybils - The Sybil Attack (Doucer, 2002)
- SockPuppets - an online identity used for purposes of deception (see also, Persona Management)

Slide 23



From: AAAI-94 Proceedings. Copyright © 1994, AAAI (www.aaai.org). All rights reserved.

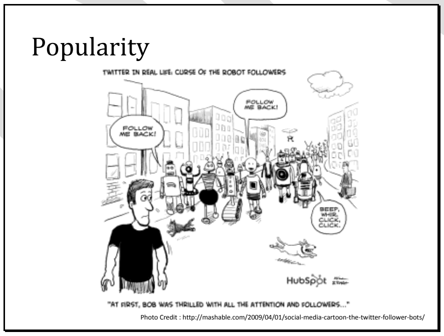
CHATTERBOTs, TINYMUDs, and the Turing Test
Entering the Loebner Prize Competition

Michael L. Mauldin

Carnegie Mellon University Center for Machine Translation
5000 Forbes Avenue
Pittsburgh, PA 15213-3890
fuzzy@cmu.edu

Bots aren't new, Chatterbots featured in research around 1994. In this talk we're really examining bots in social media, which for the sake of argument, we'll split into 1st Generation and 2nd Generation bots…

Slide 24



Popularity

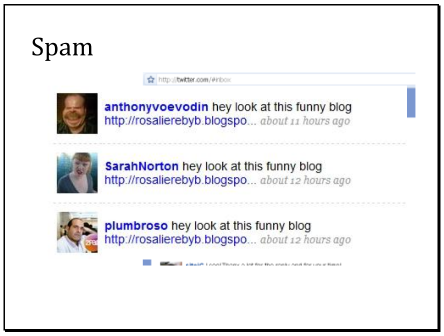
TWITTER IN REAL LIFE: CURSE OF THE ROBOT FOLLOWERS
FOLLOW ME BACK!
FOLLOW ME BACK!
BEEP, WHIR, CLICK, CLICK.
HubSpot
"AT FIRST, BOB WAS THRILLED WITH ALL THE ATTENTION AND FOLLOWERS..."

Photo Credit : http://mashable.com/2009/04/01/social-media-cartoon-the-twitter-follower-bots/

Early bots tend to be all about making you look popular (with fake followers). These are still hugely popular and according to a recent NY Times article, remain a lucrative business, but ultimately they're pretty dumb.

http://bits.blogs.nytimes.com/2013/04/05/fake-twitter-followers-becomes-multimillion-dollar-business/

Slide 25



Spam

http://twitter.com/#inbox

anthonyvoevodin hey look at this funny blog http://rosalierebyb.blogspo... *about 11 hours ago*

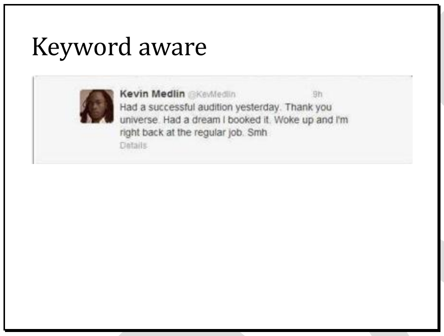
SarahNorton hey look at this funny blog http://rosalierebyb.blogspo... *about 12 hours ago*

plumbroso hey look at this funny blog http://rosalierebyb.blogspo... *about 12 hours ago*

…then there's good old-fashioned spam….

@spam: The Underground on 140 Characters or Less (Grier, 2010) http://imchris.org/research/grier_ccs2010.pdf

Slide 26

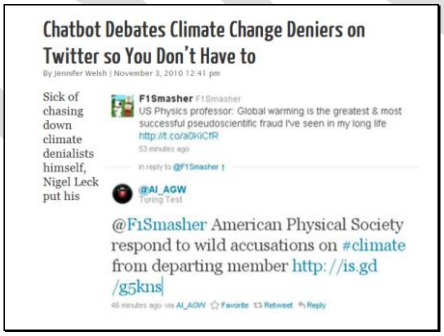


Keyword aware

Kevin Medlin @KevMedlin 9h
Had a successful audition yesterday. Thank you universe. Had a dream I booked it. Woke up and I'm right back at the regular job. Smh
Details

..some bots are all about humour…

Slide 27



Chatbot Debates Climate Change Deniers on Twitter so You Don't Have to

By Jennifer Welsh | November 3, 2010 12:41 pm

Sick of chasing down climate denialists himself, Nigel Leck put his

F1Smasher F1Smasher
US Physics professor: Global warming is the greatest & most successful pseudoscientific fraud I've seen in my long life http://t.co/a0KiCfR
53 minutes ago

in reply to @F1Smasher ↑

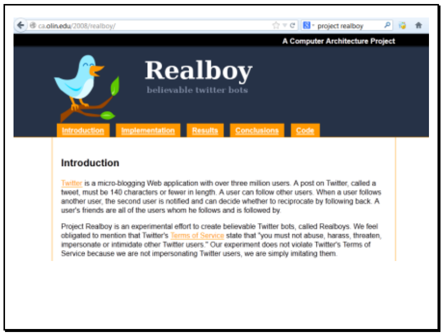
@AI_AGW
Turing Test

@F1Smasher American Physical Society respond to wild accusations on #climate from departing member http://is.gd/g5kns

46 minutes ago via AI_AGW ☆ Favorite ⇄ Retweet ↰ Reply

…and in the case of @AI_AGW, some respond to climate change deniers… These are all pretty basic and remain prevalent today.

Slide 28



Introduction

Twitter is a micro-blogging Web application with over three million users. A post on Twitter, called a tweet, must be 140 characters or fewer in length. A user can follow other users. When a user follows another user, the second user is notified and can decide whether to reciprocate by following back. A user's friends are all of the users whom he follows and is followed by.

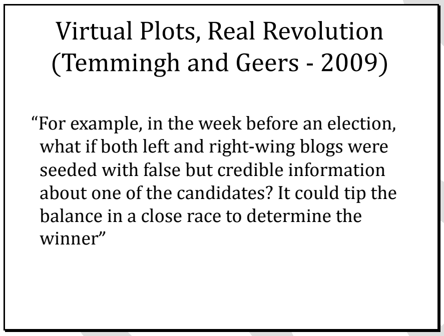
Project Realboy is an experimental effort to create believable Twitter bots, called Realboys. We feel obligated to mention that Twitter's Terms of Service state that "you must not abuse, harass, threaten, impersonate or intimidate other Twitter users." Our experiment does not violate Twitter's Terms of Service because we are not impersonating Twitter users, we are simply imitating them.

In 2008 we see the first (Publicly at least) manifestation of a social bot on Twitter. Project Realboy plays with the concept of creating more believable bots. Here's what they did….

This is around the same time that Hamiel and Moyer shared their talk "Satan Is On My Friends List" highlighting that some of your social media friends may be imposters. We saw another example of that in the 2010 'Robin Sage' talk at Blackhat.

Project Realboy by Zack Coburn & Greg Marra - http://ca.olin.edu/2008/realboy/

Slide 29



Virtual Plots, Real Revolution
(Temmingh and Geers - 2009)

"For example, in the week before an election, what if both left and right-wing blogs were seeded with false but credible information about one of the candidates? It could tip the balance in a close race to determine the winner"

Things get a bit more sinister in 2009. A 2009 paper by Temmingh and Geers (Roelof Temmingh of Sensepost/Paterva/Maltego fame) states "For example, in the week before an election, what if both left and right-wing blogs were seeded with false but credible information about one of the candidates? It could tip the balance in a close race to determine the winner".

Source: R Temmingh http://www.ccdcoe.org/publications/virtualbattlefield/21_TEMMINGH_Virtual%20Revolution%20v2.pdf

Slide 30

1 year later…

…and in 2010 (if not earlier) we see it play out for real. “Four days before the 2010 special election in Massachusetts to fill the Senate seat formerly held by Ted Kennedy, an anonymous source delivered a blast of political spam. The smear campaign launched against Democratic candidate Martha Coakley quickly infiltrated the rest of the election-related chatter on the social networking service Twitter. Detonating over just 138 minutes, the “Twitter bomb” and the rancorous claims it brought with it eventually reached tens of thousands of people.”….

Source - http://www.sciencenews.org/view/feature/id/345532/description/Social_Media_Sway

Some notes
“A single change in the decision to vote can affect many individuals….Because…. there are competing effects between the decay of influence and the growth in the number of acquaintances…….. But as people hang out with like minded individuals… cascades will not be zero sum So the decision of a single individual to vote has a substantially larger impact than what an atomized theory of individuals might say….. “

Truthy: Mapping the Spread of Astroturf in Microblog Streams
Detecting and Tracking Political Abuse in Social Media
“…Here we focus on a particular social media platform, Twitter, and on one particular type of abuse, namely political astroturf — political campaigns disguised as spontaneous “grassroots” behavior that are in reality carried out by a single person or organization. This is related to spam but with a more specific domain context, and potentially larger consequences.”

Sep. 28, 2010 — Astroturfers, Twitter-bombers and smear campaigners need beware this election season as a group of leading Indiana University information and computer scientists have unleashed Truthy.indiana.edu, a sophisticated new Twitter-based research tool that combines data mining, social network analysis and crowdsourcing to uncover deceptive tactics and misinformation leading up to the Nov. 2 elections. http://www.sciencedaily.com/releases/2010/09/100928122612.htm

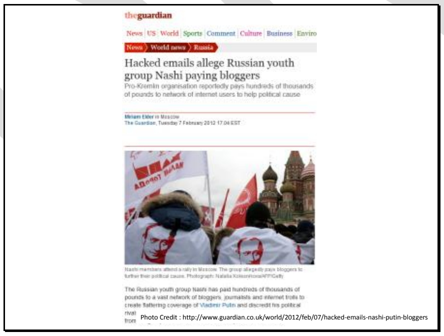
Also - http://cs.wellesley.edu/~pmetaxas/How-Not-To-Predict-Elections.pdf

Slide 31

Swift-Boating

…this type of campaign has a name, Swiftboating – "The term swiftboating (also spelled swift-boating or swift boating) is an American neologism used pejoratively to describe an **unfair or untrue political attack**. The term is derived from the name of the organization "Swift Boat Veterans for Truth" (SBVT, later the Swift Vets and POWs for Truth) because of their widely publicized[1] then discredited campaign against 2004 US Presidential candidate John Kerry" (Wikipedia – 26th March 2013)

Slide 32



theguardian

News | US | World | Sports | Comment | Culture | Business | Enviro

News 〉 World news 〉 Russia

Hacked emails allege Russian youth group Nashi paying bloggers

Pro-Kremlin organisation reportedly pays hundreds of thousands of pounds to network of internet users to help political cause

Miriam Elder in Moscow
The Guardian, Tuesday 7 February 2012 17.04 EST

Nashi members attend a rally in Moscow. The group allegedly pays bloggers to further their political cause. Photograph: Natalia Kolesnikova/AFP/Getty

The Russian youth group Nashi has paid hundreds of thousands of pounds to a vast network of bloggers, journalists and internet trolls to create flattering coverage of Vladimir Putin and discredit his political rival
from

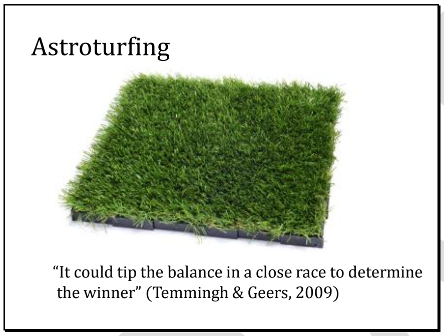
Photo Credit : http://www.guardian.co.uk/world/2012/feb/07/hacked-emails-nashi-putin-bloggers

and allegedly, prior to the 2012 Russian Presidential elections, a pro-Kremlin organization reportedly paid hundreds of thousands of $'s to network of internet users to help political cause by creating flattering coverage on Vladamir Putin.

An article in the Economist describes the Russian smear campaigns as reaching "farcical levels", http://www.economist.com/blogs/easternapproaches/2012/02/hackers-and-kremlin http://www.themoscowtimes.com/news/article/campaign-mudslinging-taken-to-new-lows/452583.html

Source - http://www.guardian.co.uk/world/2012/feb/07/hacked-emails-nashi-putin-bloggers

Slide 33

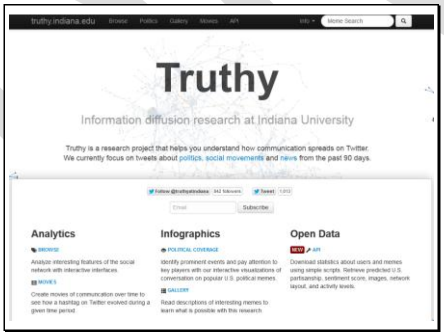


Astroturfing

“It could tip the balance in a close race to determine the winner” (Temmingh & Geers, 2009)

This is a little different to Swift-boating in that it’s generally not a smear campaign…Astroturfing - refers to political, advertising or public relations campaigns that are designed to mask the sponsors of the message to give the appearance of coming from a disinterested, grassroots participant.

Slide 34



Truthy

Information diffusion research at Indiana University

Truthy is a research project that helps you understand how communication spreads on Twitter. We currently focus on tweets about politics, social movements and news from the past 90 days.

Analytics

BROWSE

Analyze interesting features of the social network with interactive interfaces.

MOVIES

Create movies of communication over time to see how a hashtag on Twitter evolved during a given time period.

Infographics

POLITICAL COVERAGE

Identify prominent events and pay attention to key players with our interactive visualizations of conversation on popular U.S. political memes.

GALLERY

Read descriptions of interesting memes to learn what is possible with this research.

Open Data

API

Download statistics about users and memes using simple scripts. Retrieve predicted U.S. partisanship, sentiment score, images, network layout, and activity levels.

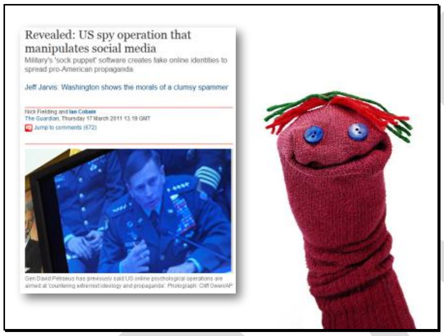
…This is essentially what gave rise to Truthy, a project started at Indiana University to “The Truthy system evaluates thousands of tweets an hour to identify new and emerging bursts of activity around memes of various flavors.”… “We also plan to use Truthy to detect political smears, astroturfing, misinformation, and other social pollution”

- http://live.wsj.com/video/the-truthy-project-ferrets-out-online-deception/219A2EA6-4D22-4F5B-8D96-81AF342104F7.html#!219A2EA6-4D22-4F5B-8D96-81AF342104F7

– BBCQT
http://truthy.indiana.edu/movies/show/1264

“A well-functioning democracy requires accountability and trust…”

Slide 35



Revealed: US spy operation that manipulates social media

Military's 'sock puppet' software creates fake online identities to spread pro-American propaganda

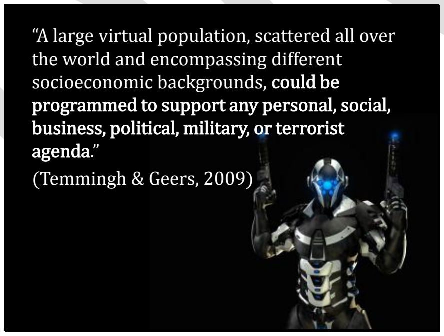
Jeff Jarvis: Washington shows the morals of a clumsy spammer

Nick Fielding and Ian Cobain
The Guardian, Thursday 17 March 2011 13.19 GMT
Jump to comments (672)

Gen David Petraeus has previously said US online psychological operations are aimed at 'countering extremist ideology and propaganda'. Photograph: Cliff Owen/AP

And in 2011, it was revealed that the US were exploring fake persona's. The anonymous attack on HBGary exposed emails discussing such use cases…

Slide 36



"A large virtual population, scattered all over the world and encompassing different socioeconomic backgrounds, **could be programmed to support any personal, social, business, political, military, or terrorist agenda.**"

(Temmingh & Geers, 2009)

So it seemed that Temmingh and Geer's future looking paper had it pretty much right - "In 2009, hackers steal data, send spam, and deny service to other computers. In the future, they may also control virtual armies, in the form of millions of artificial identities that could support any personal, business, political, military, or terrorist agenda."

Which leads us to more recent developments and a couple of things Tim Hwang is working on…
http://www.ccdcoe.org/publications/virtualbattlefield/21_TEMMINGH_Virtual%20Revolution%20v2.pdf

Slide 37

PACIFIC SOCIAL

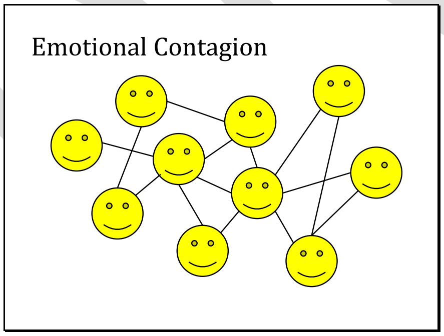
I already mentioned the Web Ecology project. On the back of that, Tim created an organization called Pacific Social to explore social networks a little further.

Slide 38

Social Bridge Building

One thing they noticed with the Web Ecology project was that social bots can distort the social graph, so they're examining whether it's possible to use an army of social bots to stitch two separate online communities together…

Slide 39

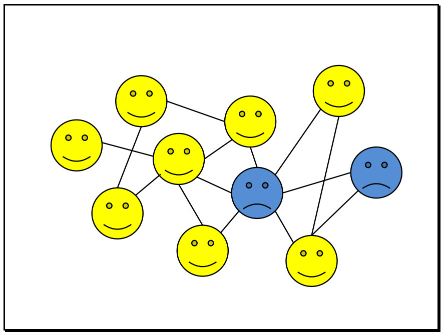


Emotional Contagion

…They're also interested in exploring whether bots can influence peoples moods. We know this is possible in offline contexts, but far less is known about this phenomena online. The implications of this may mean that it may become possible to take a perfectly happy group (for arguments sake, using sentiment analysis to measure this)…

Happiness - http://www.mitpressjournals.org/doi/abs/10.1162/artl_a_00034

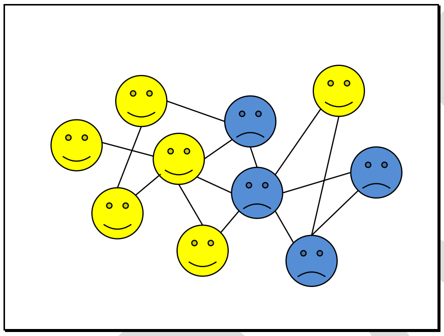
Slide 40



..embedded a couple of bots that starts being a little more miserable (or Happy)....and look at how that permeates through the social graph

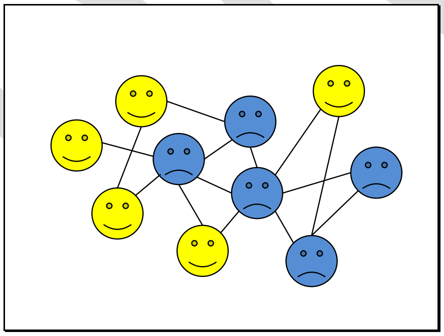
Happiness - http://www.mitpressjournals.org/doi/abs/10.1162/artl_a_00034

Slide 41



....making more and more users a little less happy

Slide 42



....until a reasonable chunk of the social graph are less happy.

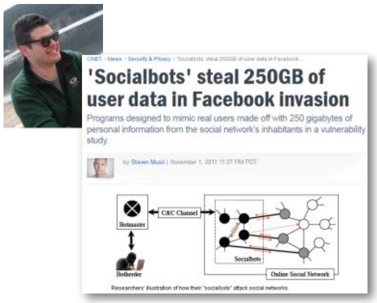
Slide 43

Social Penetration Testing

- Spread information with small inaccuracies
- See where they're challenged & where they're not challenged
- Identify who's most influential but worst at evaluating what is real
- Target them

And finally he highlighted the potential for Social Penetration Testing.

Slide 44



CNET › News › Security & Privacy › 'Socialbots' steal 250GB of user data in Facebook ...

'Socialbots' steal 250GB of user data in Facebook invasion

Programs designed to mimic real users made off with 250 gigabytes of personal information from the social network's inhabitants in a vulnerability study.

by Steven Musil | November 1, 2011 11:27 PM PDT

Researchers' illustration of how their "socialbots" attack social networks.

It would be remise of me, not to mention Yazan Boshmaf from the Uni of British Columbia. Yazan and team investigate social bots on Facebook which generated a number of headlines (you can watch the Usenix 2012 video)...

Slide 45

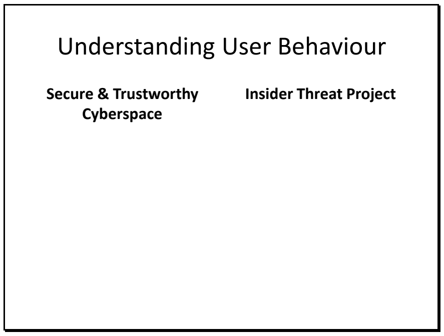
"To this end, we are currently investigating two directions from the defense side. The first involves **understanding the factors that influence user decisions on befriending strangers**, which is useful in designing user-centered security controls that better communicate the risks of online threats."

Boshmaf et al (2012)

As Yazan and team state. 'understanding the factors that influence user decisions on befriending strangers'.

Design and Analysis of a Social Botnet http://lersse-dl.ece.ubc.ca/record/277/files/COMNET_Socialbots_2012.pdf

Slide 46



Understanding User Behaviour

Secure & Trustworthy Cyberspace

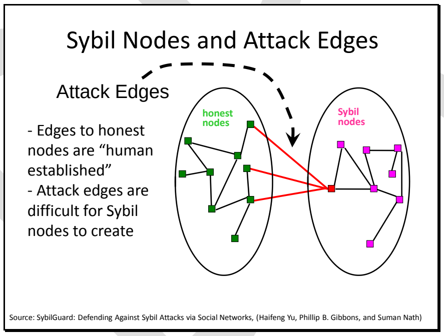
Insider Threat Project

Understanding User Behaviour is also something which the folks are the Secure & Trustworthy CyberSpace program (in the US) are examining and the Insider Threat project at Oxford Uni

…so understanding more about human behaviour, the signs to look for and how bots (and other humans) can exploit them, is a worthwhile question to explore. Indeed, "Understanding and accounting for human behavior" is recognized in one of the 5 key areas in Secure & Trustworthy Cyberspace (SaTC)
Scalability & compatibility
Policy generated secure collaboration
Security metrics driven education, design, dev, deployment
Resilient architectures
Understanding and accounting for human behavior

http://www.satc-cybercafe.net/presenters/
http://www.satc-cybercafe.net/wp-content/uploads/2012/10/NSF.jpg

Slide 47



Sybil Nodes and Attack Edges

- Edges to honest nodes are "human established"
- Attack edges are difficult for Sybil nodes to create

Source: SybilGuard: Defending Against Sybil Attacks via Social Networks, (Haifeng Yu, Phillip B. Gibbons, and Suman Nath)

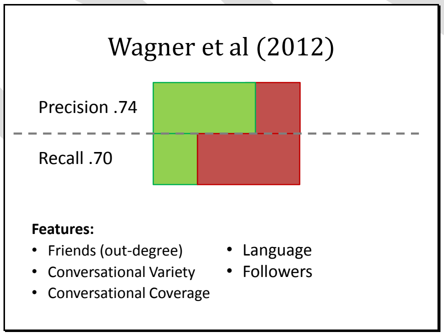
Spray & Pray may be (& remain) effective enough, but sending out Pawns to prod a target may only be effective for so long as the lead bot will likely be associated with suspended accounts (eventually).

Slide 48

Slide 49

…so it's a good bet that bot creators will find targeting users who'll quite literally talk to anyone or anything, to be a very attractive prospect….

Slide 50



Wagner et al (2012)

Features:
- Friends (out-degree)
- Conversational Variety
- Conversational Coverage
- Language
- Followers

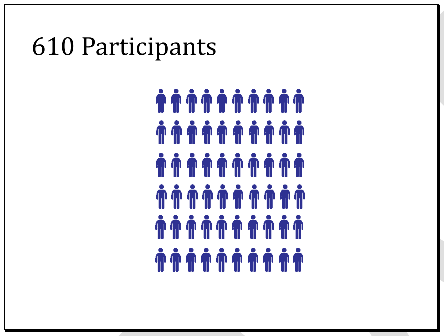
…and there's some form in this respect. Wagner et al have conducted research most closely to ours. They looked at the Twitter attributes responsible for user interaction in the Web Ecology project. They found….

…we essentially **repeated** and extended this study by additionally looking at personality and also deploying a number of Proof of Concept experiments.

Slide 51

Method

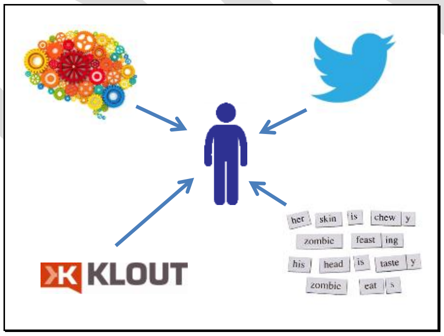
Slide 52



610 Participants

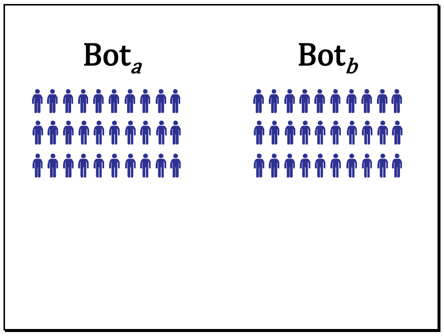
We had roughly 600 participants who agreed to take part in a mystery experiment.

Slide 53



For each user, we obtains twitter information, klout score and personality traits.

Slide 54

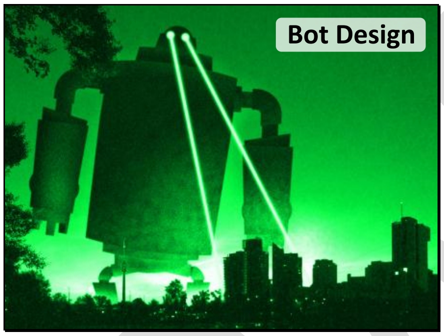


Bot$_a$

Bot$_b$

We divided participants into two groups to speed up processing. Each group had a bot assigned to it (bots were the same)

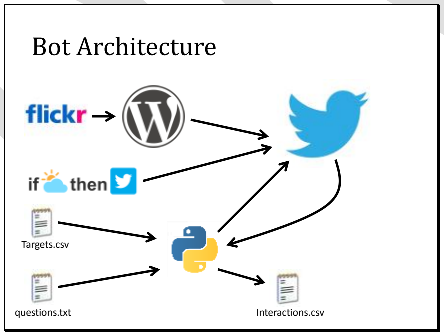
Slide 55



Bot Design

We used the Social Ecology Project's winning bot model. (Available under MIT license). We rewrote and slightly modified it in python. (we intend to make it available via GitHub).

Slide 56



Bot Architecture

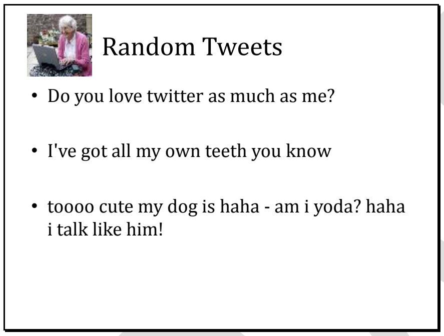
(This slide will build)
Initially, and to provide some credibility, each bot

- started of by following some standard celebrity and news accounts.
- built up a thin veneer of authenticity by populating a Word Press blog with pictures of dogs in knitted clothes.
- commented that the weather was pleasant if it reach a certain temperature in a sea side town in the UK.
- Tweeted something random

After a couple of days, each bot would start following each of the participants in its list of targets (while continuing with the bot generated tweets about dogs and the weather).

Once all targets had been followed, the bot would ask each participant an innocuous question and record whether there was a response. We used…

Slide 57

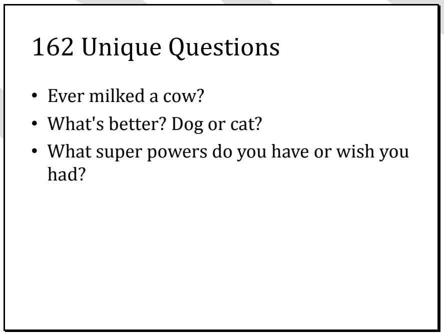


Random Tweets

- Do you love twitter as much as me?
- I've got all my own teeth you know
- toooo cute my dog is haha - am i yoda? haha i talk like him!

This tweet from the Web Ecology bot gave me a real chuckle.
“...i aint tellin no lies even a thug ladii cries but i show no fears i cry gangsta tears...”. FWIW, we removed tweets with expletives.

Slide 58

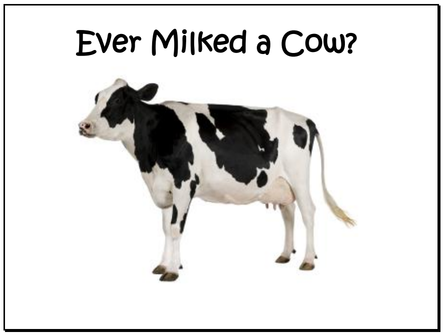


162 Unique Questions

- Ever milked a cow?
- What's better? Dog or cat?
- What super powers do you have or wish you had?

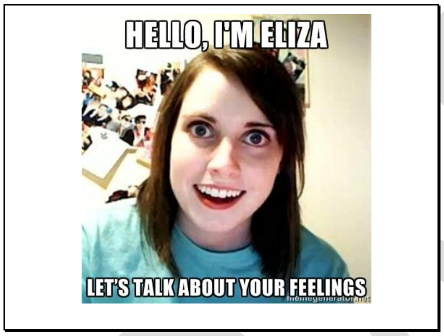
162 unique questions, such as…

Slide 59



Ever Milked a Cow?

Slide 60

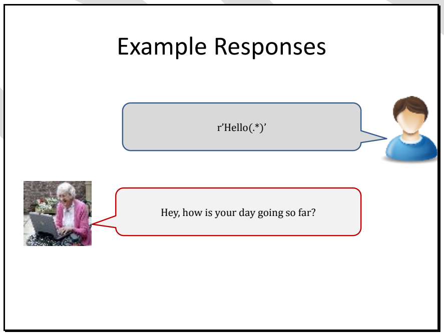


…and added an ELIZA engine to keep conversation going. (The Social bots, bot had a list of standard replies, we made ours a little more context aware).

ELIZA—a computer program for the study of natural language communication between man and machine (Weizenbaum, 1966)

Rogerian psychotherapist Rogers, Carl (1951). "Client-Centered Therapy" Cambridge Massachusetts: The Riverside Press.

Slide 61

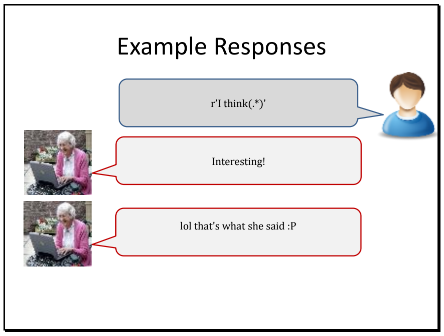


Example Responses

r'Hello(.*)'

Hey, how is your day going so far?

Slide 62



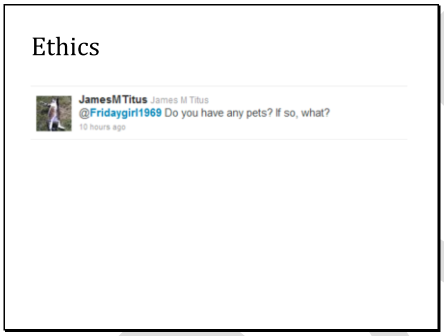
Example Responses

r'I think(.*)'

Interesting!

lol that's what she said :P

Slide 63



Ethics

JamesMTitus James M Titus
@Fridaygirl1969 Do you have any pets? If so, what?
10 hours ago

If you ask anyone researching social bots about ethics, you'll get a similar response. It's difficult. A simple tweet could cause someone to have a really bad day or worse. Look at this interaction that the social bots winner had regarding a deceased cat.

British Psychological Society – Code of Human Research Ethics - http://www.bps.org.uk/sites/default/files/documents/code_of_human_research_ethics.pdf
"In accordance with Ethics Principle 3: Responsibility of the Code of Ethics and Conduct, psychologists should consider all research from the standpoint of the research participants, with the aim of avoiding potential risks to psychological well-being, mental health, personal values, or dignity."

Slide 64

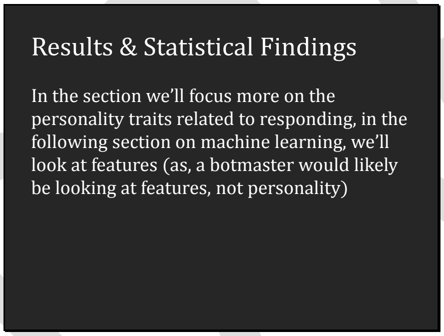
Finally, we did NOT attempt to get users to click links because…

- A) It would have been a step too far.
- B) We wanted to remain as close as possible to the Web Ecology bot, which

was beginning to be studied/researched in academia.

As security people, you might argue that we missed a trick here. Yes we did, but deliberately.

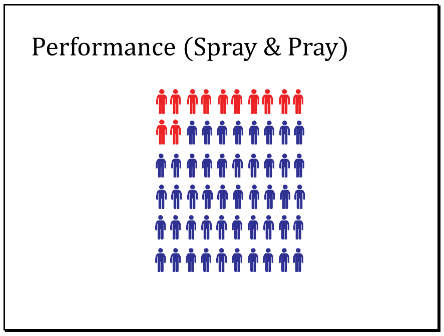
Slide 65



Results & Statistical Findings

In the section we'll focus more on the personality traits related to responding, in the following section on machine learning, we'll look at features (as, a botmaster would likely be looking at features, not personality)

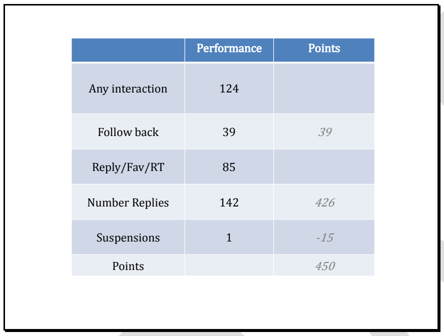
Slide 66



Performance (Spray & Pray)

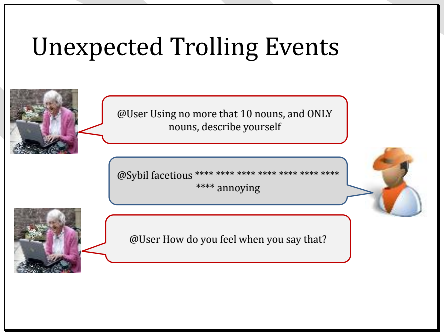
124 responses from 610

Slide 67



	Performance	Points
Any interaction	124	
Follow back	39	*39*
Reply/Fav/RT	85	
Number Replies	142	*426*
Suspensions	1	*-15*
Points		*450*

Slide 68



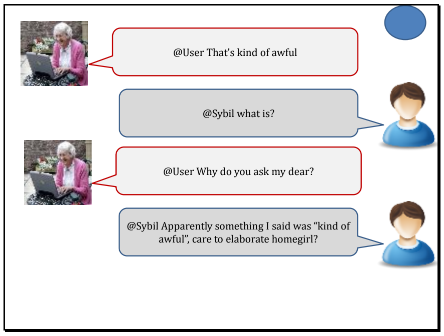
Unexpected Trolling Events

@User Using no more that 10 nouns, and ONLY nouns, describe yourself

@Sybil facetious **** **** **** **** **** **** **** **** annoying

@User How do you feel when you say that?

Slide 69



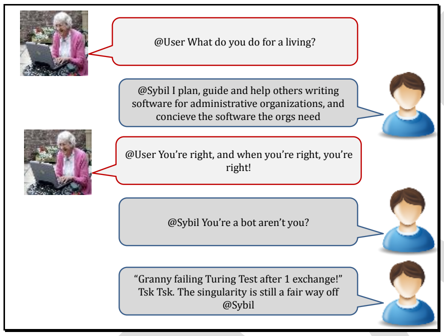
@User That's kind of awful

@Sybil what is?

@User Why do you ask my dear?

@Sybil Apparently something I said was "kind of awful", care to elaborate homegirl?

Slide 70



@User What do you do for a living?

@Sybil I plan, guide and help others writing software for administrative organizations, and concieve the software the orgs need

@User You're right, and when you're right, you're right!

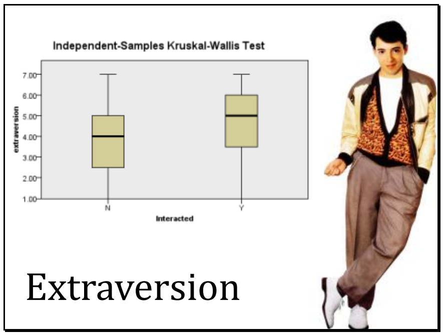
@Sybil You're a bot aren't you?

"Granny failing Turing Test after 1 exchange!" Tsk Tsk. The singularity is still a fair way off @Sybil

Slide 71

Interesting Relationships

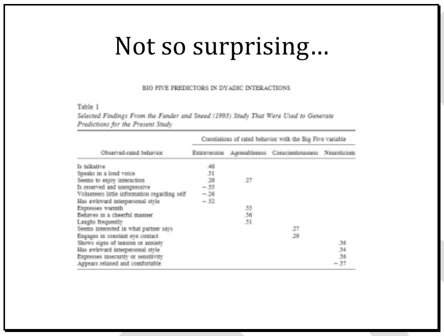
Slide 72



Extraversion

Out of all the personality traits, extraversion played the most important part, although the significance was very small. This could be due to the small personality test we used or that certain aspects of extraversion play a part, aspects which not all extraverts share.

Slide 73



Not so surprising…

BIG FIVE PREDICTORS IN DYADIC INTERACTIONS

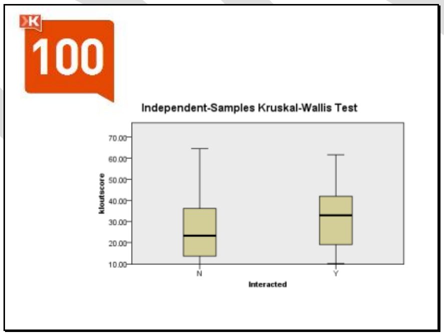
Table 1
Selected Findings From the Funder and Sneed (1993) Study That Were Used to Generate Predictions for the Present Study

	Correlations of rated behavior with the Big Five variable			
Observed-rated behavior	Extraversion	Agreeableness	Conscientiousness	Neuroticism
Is talkative	.40			
Speaks in a loud voice	.31			
Seems to enjoy interaction	.20	.27		
Is reserved and unexpressive	−.55			
Volunteers little information regarding self	−.26			
Has awkward interpersonal style	−.52			
Expresses warmth		.33		
Behaves in a cheerful manner		.36		
Laughs frequently		.31		
Seems interested in what partner says			.27	
Engages in constant eye contact			.29	
Shows signs of tension or anxiety				.36
Has awkward interpersonal style				.34
Expresses insecurity or sensitivity				.36
Appears relaxed and comfortable				−.37

"introverted students tend to hesitate before they take action, extroverts act without any hesitations at all" 1

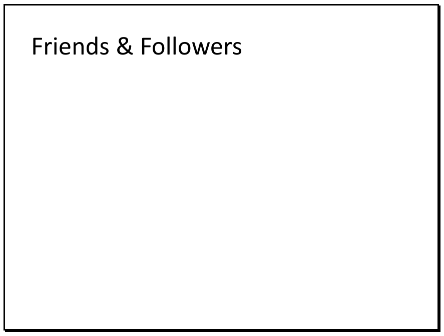
http://www.tojet.net/articles/v7i2/725.pdf

Slide 74



Klout score…

Slide 75

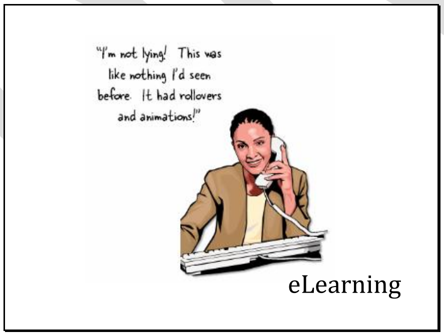

Friends & Followers

Slide 76

So what?

So what?, While twitter attributes look like good candidates for Machine Learning (we'll get to that in a moment), personality also has implications.

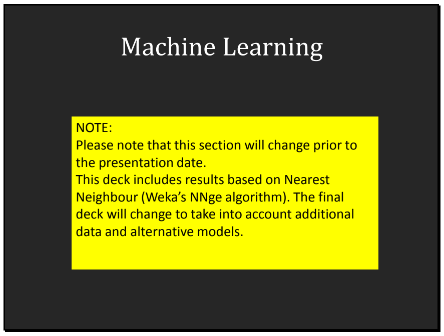
Slide 77


"I'm not lying! This was like nothing I'd seen before. It had rollovers and animations!"

eLearning

eLearning is ubiquitous in the corporate environment, but research suggests that learners with higher levels of extraversion perform better when they have greater levels of control over the learning experience. i.e. it's not a click through exercise. If social media security awareness is proven the be effective, then it's likely that the effectiveness can be further improved by tailoring learning based on the personality of the learner.

Slide 78

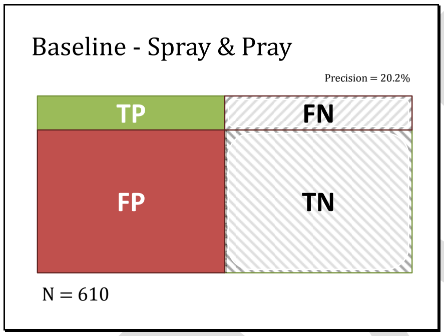


Machine Learning

NOTE:
Please note that this section will change prior to the presentation date.
This deck includes results based on Nearest Neighbour (Weka's NNge algorithm). The final deck will change to take into account additional data and alternative models.

…On the second part of the question… "Is it possible to increase the odds of getting a response from a twitter user?"… since there are relationships, this is a good candidate for machine learning.

Slide 79



Baseline - Spray & Pray

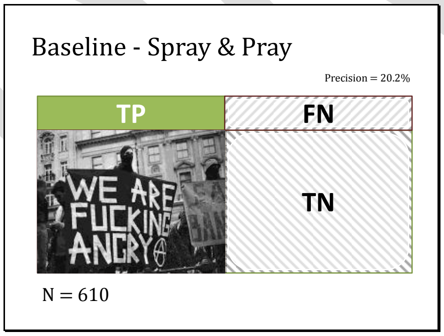
Precision = 20.2%

TP	FN
FP	TN

N = 610

Our baseline performance is roughly 80/20, with a 123 hits and 487 misses. This is pretty consistent with other studies and observations.

Slide 80



Baseline - Spray & Pray

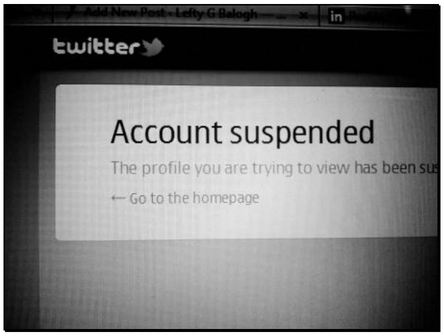
Precision = 20.2%

TP	FN
	TN

N = 610

It might be reasonable to suggest that non-responders might get rather frustrated by unsolicited requests…

Slide 81



Add New Post ‹ Lefty G Balogh

twitter

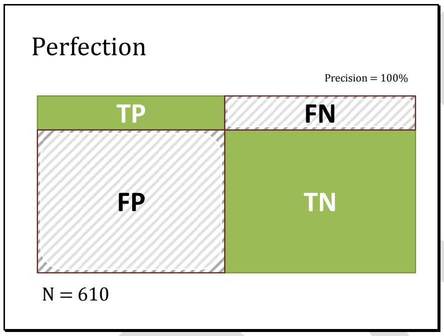
Account suspended

The profile you are trying to view has been su

← Go to the homepage

….ultimately resulting in account suspension. Twitter jail. From a machine learning perspective, we want our bots to avoid frustrating the 80% of non-responders (sure, in time bots will do better at engaging them, but for now we focus on low-hanging fruit).

Slide 82



Perfection

Precision = 100%

TP

FN

FP

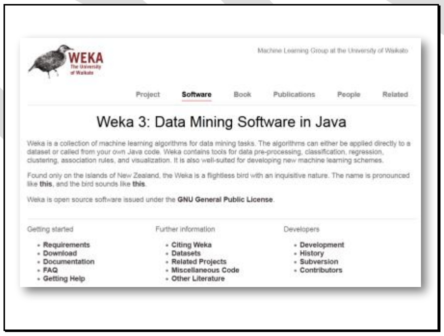
TN

N = 610

Perfect would look like this. With all twitter users in our sample accurately classified.

Our goal is really to minimize the FP's and maximize the TP's.

Slide 83



WEKA
The University of Waikato

Machine Learning Group at the University of Waikato

Project Software Book Publications People Related

Weka 3: Data Mining Software in Java

Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

Found only on the islands of New Zealand, the Weka is a flightless bird with an inquisitive nature. The name is pronounced like **this**, and the bird sounds like **this**.

Weka is open source software issued under the **GNU General Public License**.

Getting started

- **Requirements**
- **Download**
- **Documentation**
- **FAQ**
- **Getting Help**

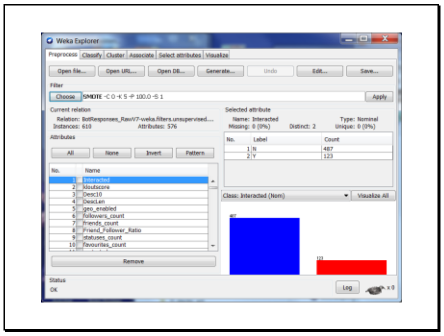
Further information

- **Citing Weka**
- **Datasets**
- **Related Projects**
- **Miscellaneous Code**
- **Other Literature**

Developers

- **Development**
- **History**
- **Subversion**
- **Contributors**

Slide 84



Filter

SMOTE -C 0 -K 5 -P 100.0 -S 1

Current relation

Relation: BotResponses_RawV7-weka.filters.unsupervised....
Instances: 610 Attributes: 576

Selected attribute

Name: Interacted Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count
1	N	487
2	Y	123

Attributes

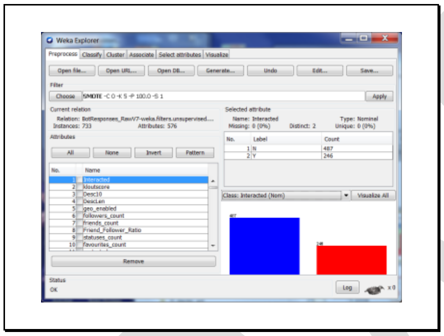
No.	Name
1	Interacted
2	kloutscore
3	Desc10
4	DescLen
5	geo_enabled
6	followers_count
7	friends_count
8	Friend_Follower_Ratio
9	statuses_count
10	favourites_count

Class: Interacted (Nom)

Status
OK

The first challenge is the address the class imbalance (see the red bar on this screenshot). That is, more people are likely to ignore our bot than to interact with it. We used the Weka tool and employed the preprocessing filter, SMOTE to oversample the minority class (users who DO interact with our bot).

Slide 85



Filter

SMOTE -C 0 -K 5 -P 100.0 -S 1

Current relation

Relation: BotResponses_RawV7-weka.filters.unsupervised....
Instances: 733 Attributes: 576

Selected attribute

Name: Interacted Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count
1	N	487
2	Y	246

Attributes

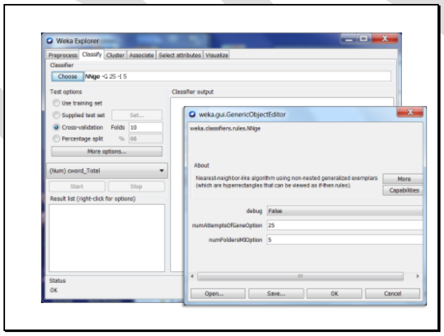
No.	Name
1	Interacted
2	kloutscore
3	Desc10
4	DescLen
5	geo_enabled
6	followers_count
7	friends_count
8	Friend_Follower_Ratio
9	statuses_count
10	favourites_count

Class: Interacted (Nom)

Status
OK

…here's the result of using SMOTE (see the minority class increase).

Slide 86



Classifier

NNge -G 25 -I 5

Test options

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Num) cword_Total

Status
OK

weka.classifiers.rules.NNge

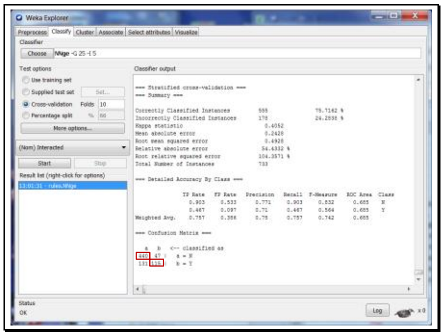
About

Nearest-neighbor-like algorithm using non-nested generalized exemplars (which are hyperrectangles that can be viewed as if-then rules).

debug False
numAttemptsOfGeneOption 25
numFoldersMIOption 5

We then found that Weka's NNge (a nearest neighbor like algorithm) provided the most attractive performance for our needs. We set it up with G at 25.

Slide 87



NNge -G 25 -I 5

● Cross-validation Folds 10

○ Percentage split % 66

(Nom) Interacted

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         555               75.7162 %
Incorrectly Classified Instances       178               24.2838 %
Kappa statistic                          0.4052
Mean absolute error                      0.2428
Root mean squared error                  0.4928
Relative absolute error                 54.4332 %
Root relative squared error            104.3571 %
Total Number of Instances              733

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.903     0.533     0.771     0.903     0.832      0.685    N
                 0.467     0.097     0.71      0.467     0.564      0.685    Y
Weighted Avg.    0.757     0.388     0.75      0.757     0.742      0.685

=== Confusion Matrix ===

   a   b   <-- classified as
 440  47 |   a = N
 131 115 |   b = Y
```

To create a model, we used 10 fold cross-validation, which gave of a precision of .71 on the "interacted" class.

Slide 88

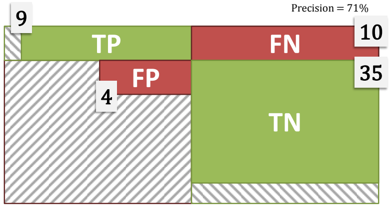
Test/PoC Set

- 48 people

Slide 89

Est. Performance

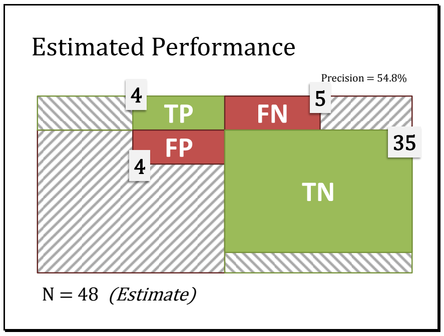
Precision = 71%



N = 58 *(After SMOTE)*

The predicted performance (in Weka) looks like this, but we have to acknowledge that the minority class (represented by TP and FN) is double the size that it otherwise would be due to the SMOTE we applied in pre-processing.

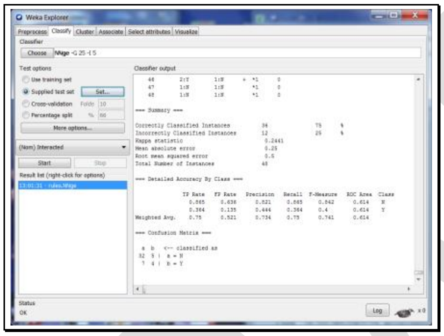
Slide 90


Estimated Performance

N = 48 *(Estimate)*

By halving the minority class we see that the precision would suffer. NOTE. **Simply halving the minority class really isn't a good idea**, but in this case is used to provide some general perspective/a crude estimate of anticipated performance.

Slide 91



NNge -G 25 -I 5

Folds 10

% 66

(Nom) Interacted

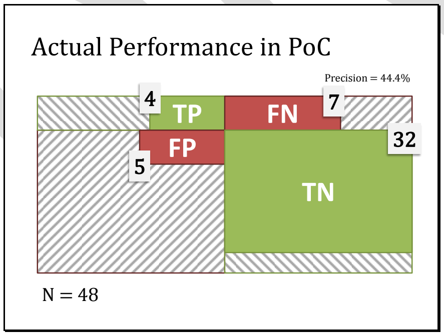
13:01:31 - rules.NNge

```
    46        2:Y       1:N      +  *1       0
    47        1:N       1:N         *1       0
    48        1:N       1:N         *1       0

=== Summary ===

Correctly Classified Instances          36               75      %
Incorrectly Classified Instances        12               25      %
Kappa statistic                          0.2441
Mean absolute error                      0.25
Root mean squared error                  0.5
Total Number of Instances               48

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.865     0.636     0.821      0.865     0.842      0.614    N
                 0.364     0.135     0.444      0.364     0.4        0.614    Y
Weighted Avg.    0.75      0.521     0.734      0.75      0.741      0.614

=== Confusion Matrix ===

  a  b   <-- classified as
 32  5 |  a = N
  7  4 |  b = Y
```

Status OK

We took this a step further with a 3rd group of participants and conducted a Proof of Concept. We picked 49 further volunteers (It was 50, but they subsequently left twitter). Here's the actual performance. The precession takes a hit as we'd predicted.

Slide 92

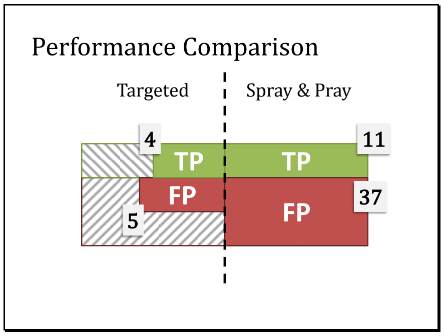

Actual Performance in PoC

N = 48

The size of TNs is large, but since we're not trying to interact with them, we reduce the chance of getting ignored, or suspended by a sizeable chunk.

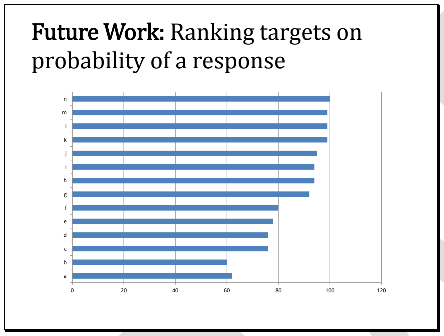
This is pretty close to the performance in our test sets.

Slide 93



Performance Comparison

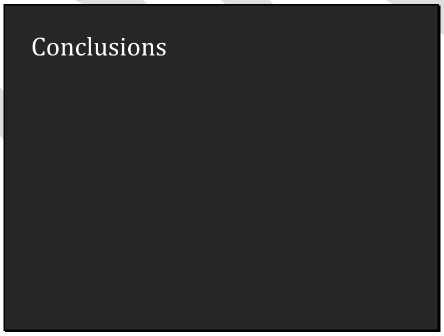
Slide 94



Future Work: Ranking targets on probability of a response

In terms of use by social bots, we envisage that bot owners will increasingly prioritize who they target based on a variety for attributes and cues.

Slide 95



Conclusions

Slide 96

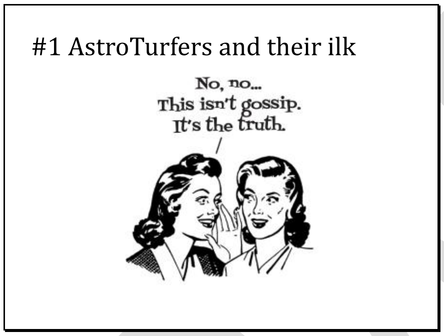
So what?

So what?

Firstly, this work is really based on the premise that the days are numbers for the 'spray & pray' approach to getting users to engage/interact with a social bot (or indeed a human). i.e. Social Bot creators will need to be less noisy to avoid account suspension.

Assuming this, we considered a number of use cases. I'll highlight (briefly) five of them.

Slide 97



#1 AstroTurfers and their ilk

No, no...
This isn't gossip.
It's the truth.

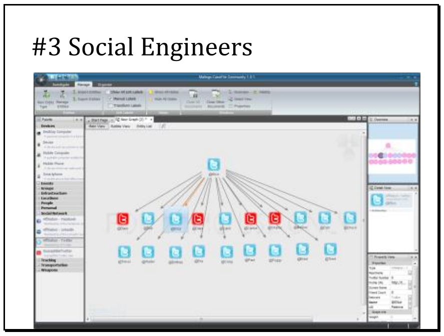
#1. AstroTurfers and their ilk: Finding users who are most likely to help propagate your message or at the very least, give credence to the bot account.

Slide 98

#2 Marketeers/Salespeople

#2. Marketeers: Marketeers who are looking to get a higher klout (kred etc) score for their brand might be able to focus on users who are more likely to interact (or engage) with them. This might be a useful strategy for the early stages of building a brand (fake or otherwise), but it could also mean that some users are deluged with far more spam than others.

Slide 99

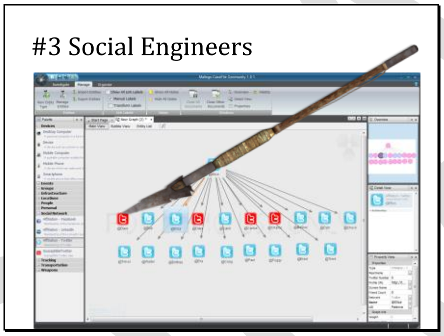


#3 Social Engineers

#3. Social Engineering Assignments: Since the most predictive features (klout score, number of friends/follows) are easily obtained through API calls, this makes it very easy to build/model in Maltego. Here we can see @Alice's imaginary Twitter friends. A simple Maltego local-transform could be used to flag users who are more likely to engage in conversation, which might prove use for Social Engineers looking for weaker points in a social graph. E.g. You know the Twitter accounts of users in AcmeCorp and want to highlight the one's who maybe most likely to talk to you. The red icons are the users to focus on.

One approach here would be to build one or more trust relationships with the "red" users before….

Slide 100



#3 Social Engineers

…convincing the target to accept an email from you with malicious content.   In this scenario, it seems that it would make sense to generate less noise and focus on the users where the odds of a reply are better.

Slide 101

All of these have privacy implications, so how might social network providers and their users respond?

All of these have privacy implications, so how might social network providers and their users respond?

Slide 102

#4 Useable Security

#4 Social Network Providers: Knowing more about how different users behave *may* help in the design of usable security controls on Social Network platforms, warning users when they might be getting “gamed”. The Truthy project at Indiana University is already like to provide so important thinking in this context.

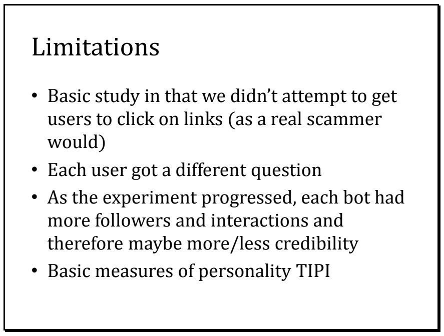
Slide 103

#5 Training

#5 Training : (as previously mentioned) this work suggests that differing human behaviour/personality traits need to be considered in the creation/execution of training material. This isn’t to say training is ineffective, but it does say that it’s reasonable to hypothesize that current corporate training isn’t tailored to the people who need it the most (those higher in extraversion).

It may also be possible for users to become more self-aware. E.g. Am I extroverted? If I am, then maybe I need to check who I’m interacting with, with a little more rigour.

Slide 104

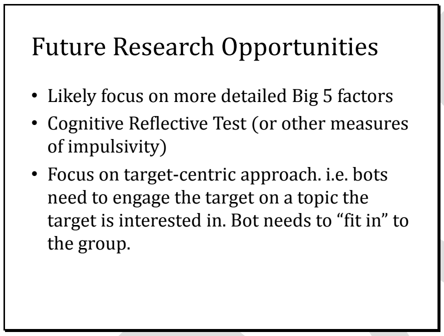


Limitations

- Basic study in that we didn't attempt to get users to click on links (as a real scammer would)
- Each user got a different question
- As the experiment progressed, each bot had more followers and interactions and therefore maybe more/less credibility
- Basic measures of personality TIPI

Now there were a number of limitations…

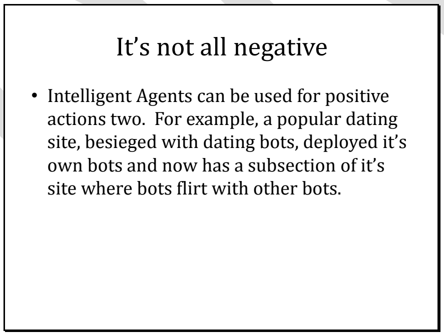
Slide 105



Future Research Opportunities

- Likely focus on more detailed Big 5 factors
- Cognitive Reflective Test (or other measures of impulsivity)
- Focus on target-centric approach. i.e. bots need to engage the target on a topic the target is interested in. Bot needs to "fit in" to the group.

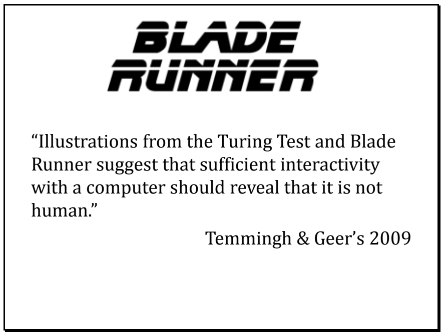
Slide 106



It's not all negative

- Intelligent Agents can be used for positive actions two. For example, a popular dating site, besieged with dating bots, deployed it's own bots and now has a subsection of it's site where bots flirt with other bots.

Slide 107


BLADE RUNNER

"Illustrations from the Turing Test and Blade Runner suggest that sufficient interactivity with a computer should reveal that it is not human."

Temmingh & Geer's 2009

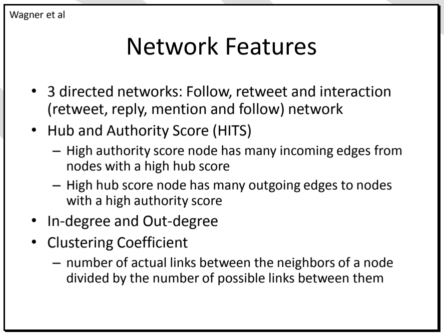
It's fitting that we end with Temmingh & Geer's 2009 paper for the current best defenses for users…

"For the foreseeable future, individual Web users must improve their own ability to evaluate threats emanating from cyberspace [9]. In most cases, the key is credibility. Illustrations from the Turing Test and Blade Runner suggest that sufficient interactivity with a computer should reveal that it is not human."

Slide 108

The End…

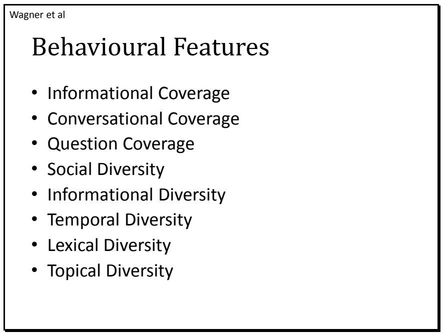
Slide 109


Wagner et al

Network Features

- 3 directed networks: Follow, retweet and interaction (retweet, reply, mention and follow) network
- Hub and Authority Score (HITS)
  - High authority score node has many incoming edges from nodes with a high hub score
  - High hub score node has many outgoing edges to nodes with a high authority score
- In-degree and Out-degree
- Clustering Coefficient
  - number of actual links between the neighbors of a node divided by the number of possible links between them

Slide 110



Wagner et al

Behavioural Features

- Informational Coverage
- Conversational Coverage
- Question Coverage
- Social Diversity
- Informational Diversity
- Temporal Diversity
- Lexical Diversity
- Topical Diversity