

Detecting Spammers on Social Networks

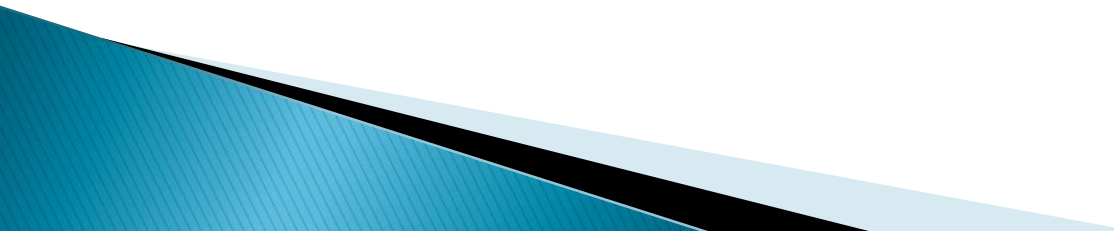
Gianluca Stringhini, Christopher Kruegel, Giovanni Vigna
University of California, Santa Barbara
26th ACSAC(December, 2010)

Presented by
Alankrit Chona 2009CS10176
Nikita Gupta 2009CS50248

Outline

- ▶ Motivation
- ▶ The Popular Social Networks
- ▶ Data Collection
- ▶ Analysis of Collected Data
- ▶ Spam Profile Detection
- ▶ Our Proposals

Idea

- ▶ To conduct a comprehensive survey of existing and potential attack behaviors in social network sites
 - ▶ Identify patterns in such attack behaviors
 - ▶ Detect Spammers in a real-world social network
 - ▶ Review existing solutions, measurement as well as defense mechanisms
- 

What is SPAM?

- ▶ “Unsolicited usually commercial e-mail sent to a large number of addresses” – Merriam Webster Online
- ▶ As the Internet has supported new applications, many other forms are common, requiring a much broader definition

Capturing user attention unjustifiably in Internet enabled applications (e-mail, Web, Social Media etc..)

SPAM TAXONOMY

INTERNET SPAM

[Forms] **DIRECT**

INDIRECT

E-Mail Spam

IM Spam (SPIM)

Social Network Spam

Spam Blogs (Splogs)

General Web Spam

[Mechanisms] Social Media Spam

Spamdexing

[Mechanisms]

NIST Computer Security Division's CSRC Home page

CSD publications, events, cryptographic standards and applications. Information on security testing, security management, and research initiatives. Includes links to the national vulnerability cyber

Source: csrc.nist.gov

Macromedia - Security Resource Center

**Auto-generated and/or
Plagiarized Content**

certain sections of

ail Security

Test / Event Log

Scan your system for trojans using this free online trojan scanner. Anti trojan software will allow Security Tests Web Site Security Audit. Get a free audit of your Website Security and check if
Source: www.windowsecurity.com

Security Center - PayPal

Welcome to the PayPal Security Center. Here, you'll find the latest information on how to buy and sell safely online. You'll get tools to help keep you protected.
Source: www.paypal.com

ASG Security - Home Security Solutions

We all know the feeling, that nagging concern in the back of our minds What if there's a fire? What if someone breaks in? ASG has a solution for you and your family - to help stop the worrying.
Source: www.asgsecurity.com

Yahoo! Security Center | About Passwords

Passwords. Your password is more than just a key to your online account. If your password falls into the wrong hands, someone

www.Protect-My-Home.com

Security System Reviewed

Top Rated Security
Systems Ratings -
Security Systems
www.a-SecuritySystems.com

**Advertisements in
Profitable Contexts**

On

Pro

Pro

Sig

Detec

Ships Today

www.NeedDecals.com

Advertise on this site

EPOXY GARAGE FLOOR
BIKE RACKS
BERNINA SEWING MACHINES
PORTABLE AIR CONDITIONERS
DISCOUNT FURNITURE STORES
ELECTRIC FIREPLACE
MASSAGE CHAIRS
FACTORY FURNITURE
PRINTABLE GROCERY COUPONS
LAND IN NORTH CAROLINA
MOEN FAUCETS
DINING FURNITURE
FIBER GLASS
BOYNTON BEACH REAL ESTATE
CHENILLE BEDSPREADS
HOUSE DESIGNS

**Link Farms to promote
other spam pages**

FOR MORE FIRST EDITION
SELL THESE

Spam on Social Networking Sites

- ▶ Sites such as *Facebook*, *MySpace*, and *Twitter* are consistently among the **top 20** most-viewed web sites of the Internet
- ▶ In 2008, **83%** of the users of social networks have received at least one **unwanted friend request** or **message**
- ▶ A previous study showed that **45%** of users on a social networking site readily **click on links** posted by their “friend” accounts, **even if they do not know** that person in real life

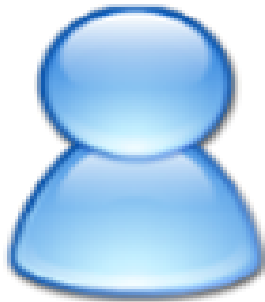


The Popular Social Networks



- Facebook administrators claim to have more than 400 million active users all over the world
- Most Facebook users were grouped in networks, where people coming from a certain country, town, or school could find their neighbours or peers
- Default privacy setting was to allow all people in the **same network** to view each other's profiles
 - Facebook deprecated geographic networks in October 2009

facebook.



You

facebook®

Your wall



You



They afsana check this embarrassing video Bizarre .



[VIDEO] Yeah!! It happens on Live Television!
play-all-now.blogspot.com

LOL Check out this video its a very embarrassing moment for her This shocking ...

That last afsana check the sad post I dare you can watch this .



[Video] Girl killed herself after her dad posted a
secret of her on here fb wall!!!
smdpskdjls.blogspot.com

click here to see dad post and emma suckle letter , you will
really be shocked... !! This cool ...



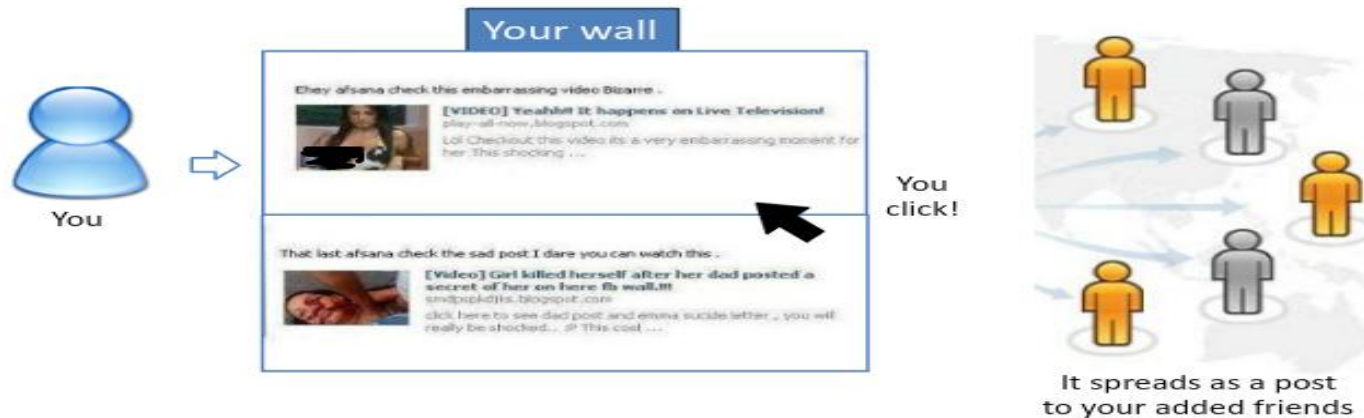
**You
CLICK !**

facebook®



It spreads as a post
to your added friends

facebook



An extension to our web browser which accesses the websites we visit and posts such content on our behalf

YOUTUBE EXTENSION

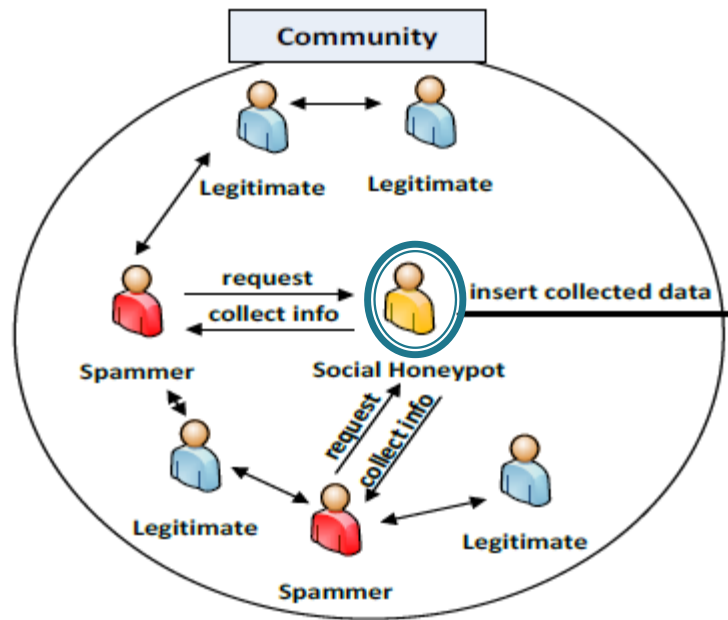
YOUTUBE PREMIUM

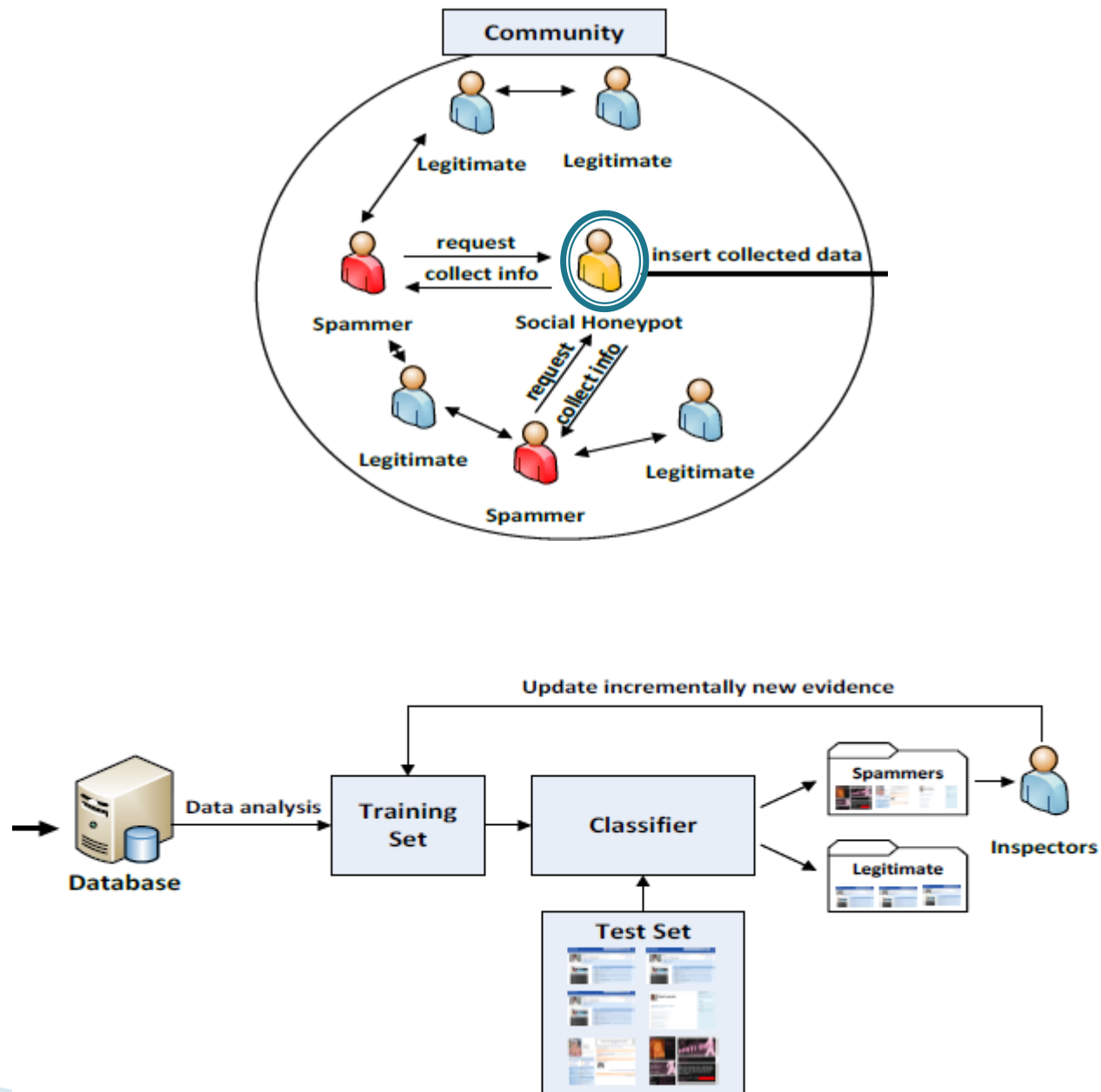


- ▶ It is designed as a microblogging platform, where users send short text messages (i.e., *tweets*) that appear on their friends' pages
- ▶ Users are identified only by a username and, optionally, by a real name
- ▶ Tweets can be grouped by hashtags, which are popular words, beginning with a “#” character
 - This allows users to efficiently search who is posting topics of interest at a certain time

Data Collection

- ▶ **900** profiles on Facebook, MySpace, and Twitter, 300 created on each platform
- ▶ A **honeypot** is a "trap" set to detect, deflect, or in some manner counteract attempts at unauthorized use of information systems
- ▶ Due to the similarity of these profiles to honeypots, these accounts are called as *honey-profiles*






Honey-Profiles

Mail Center Friend Request Manager

 Approve or Deny Your Friend Requests [[Help](#)]

Listing 1-1 of 1

1 of 1

<input type="checkbox"/>	Date:	From:	Confirmation:
<input type="checkbox"/>	9 Dec 2007 11:46 AM	  Online Now!	Delia wants to be your friend! <input type="button" value="✓ Approve"/> <input type="button" value="✗ Deny"/> <input type="button" value="✉ Spam"/> Send Message

Listing 1-1 of 1

1 of 1

☐ Check/Uncheck All

Honey-Profiles

▶ Facebook

- Joined **16 geographic networks**, using a small number of manually-created accounts
- Crawled **2,000** accounts at random, logging names, ages, and gender for each network
- Randomly mixed this data (names, surnames, and ages) and created the honey-profiles

Honey-Profiles

- ▶ MySpace
 - Crawled 4,000 accounts in total
- ▶ Twitter
 - Only information required for signing up is a full name and a profile name

Collection of Data

- ▶ No friend requests were sent, but all those that were received were accepted
- ▶ Logged every email notification received from the social networks, as well as all the requests and messages seen on the honey-profiles
- ▶ Scripts ran continuously for 12 months for Facebook (from June 6, 2009 to June 6, 2010), and for 11 months for MySpace and Twitter (from June 24, 2009 to June 6, 2010)

Collected Data

Network	Overall	Spammers
Facebook	3,831	173
MySpace	22	8
Twitter	397	361

Looking for
“local”
friends

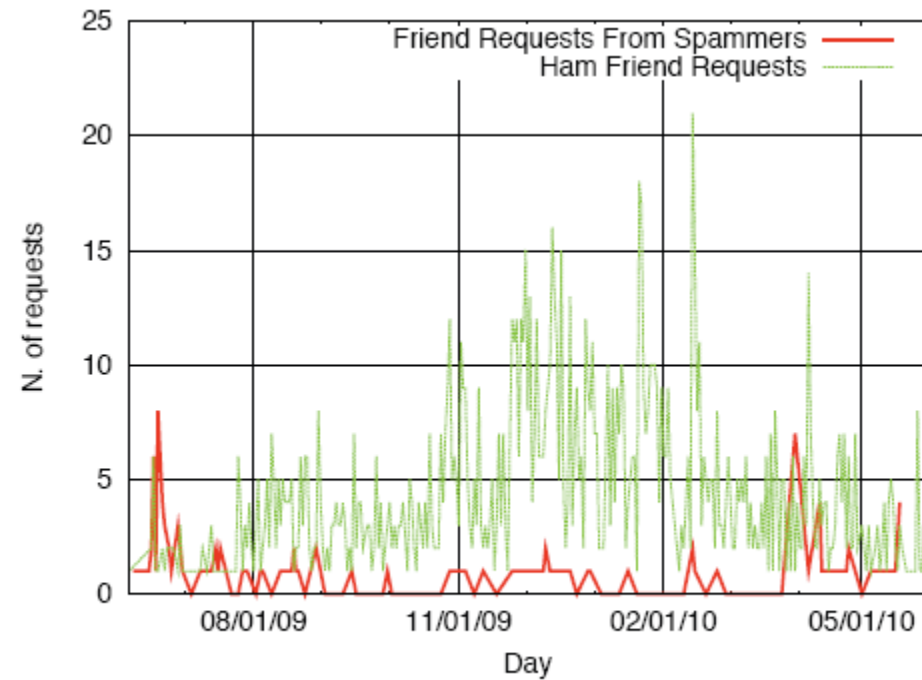
Table 1: Friend requests received on the various social networks.

Network	Overall	Spammers
Facebook	72,431	3,882
MySpace	25	0
Twitter	13,113	11,338

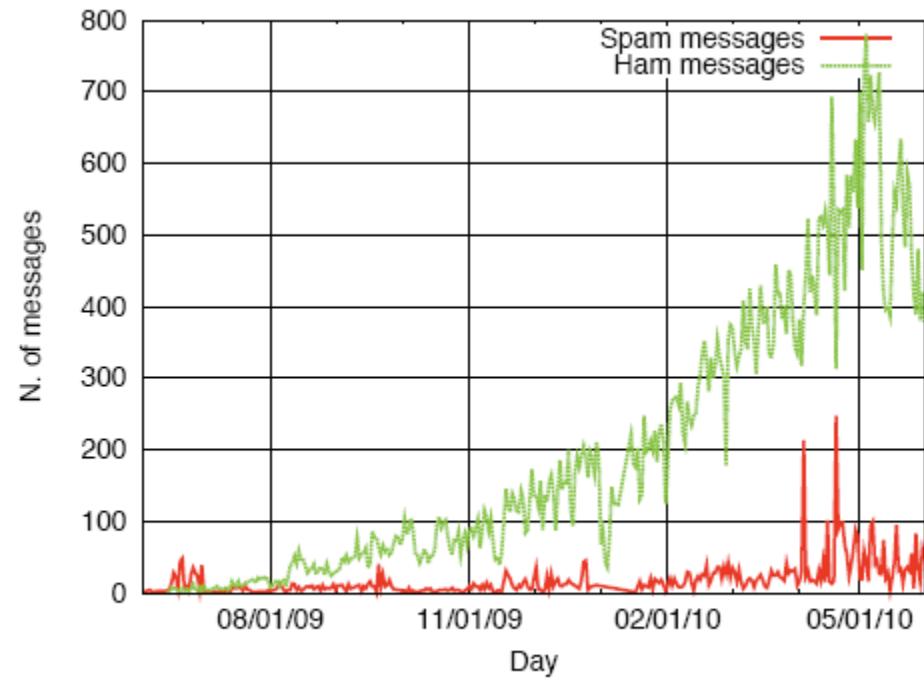
Table 2: Messages received on the various social networks.

Analysis of Collected Data

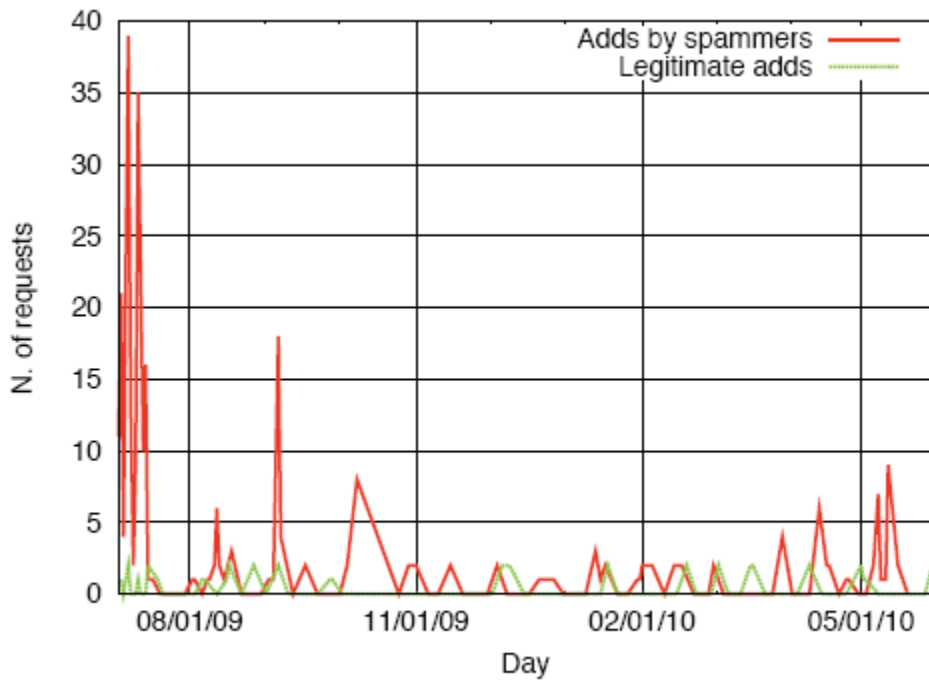
»» Facebook
Twitter



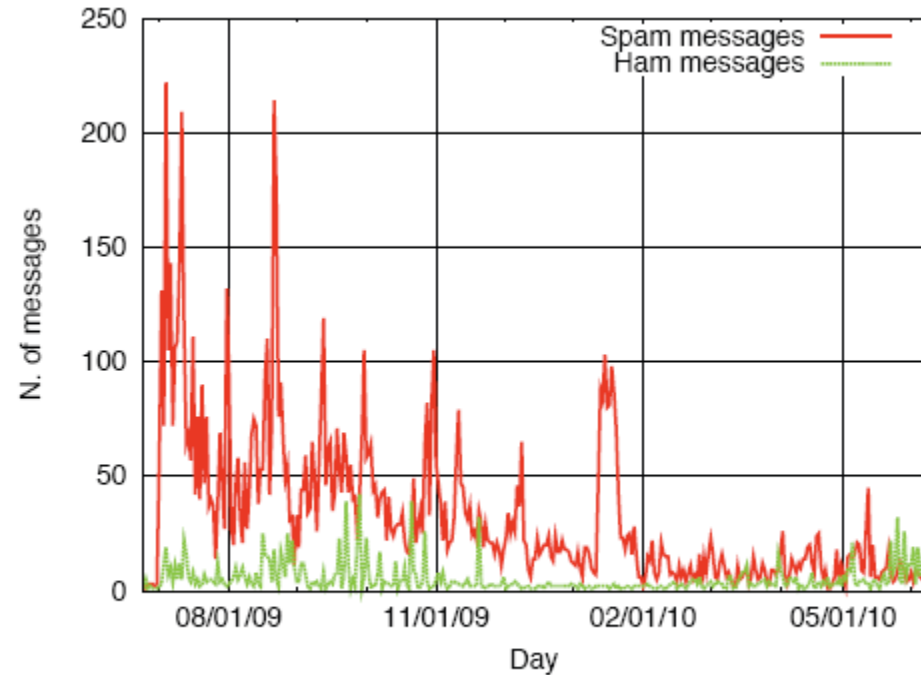
(a) Friend requests received.



(b) Messages received.



(a) Users starting following honey-profiles



(b) Messages received

Results

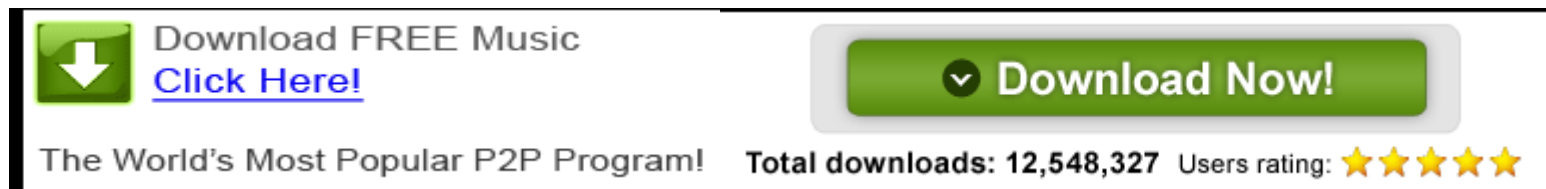
- ▶ Honey-profiles did not only receive friend requests and messages from spammers, but also a surprising amount from legitimate accounts.
 - In particular, many social network users aim to increase their popularity by adding as friends people they do not know.
- ▶ On Facebook, since all honey-profiles were members of a geographic network it may be that people looking for local “friends” would have contacted some of our accounts.

Spam Bot Analysis

»» Displayer
Bragger
Poster
Whisperer

Displayer

- ▶ Bots that only display some spam content on their own profile pages.

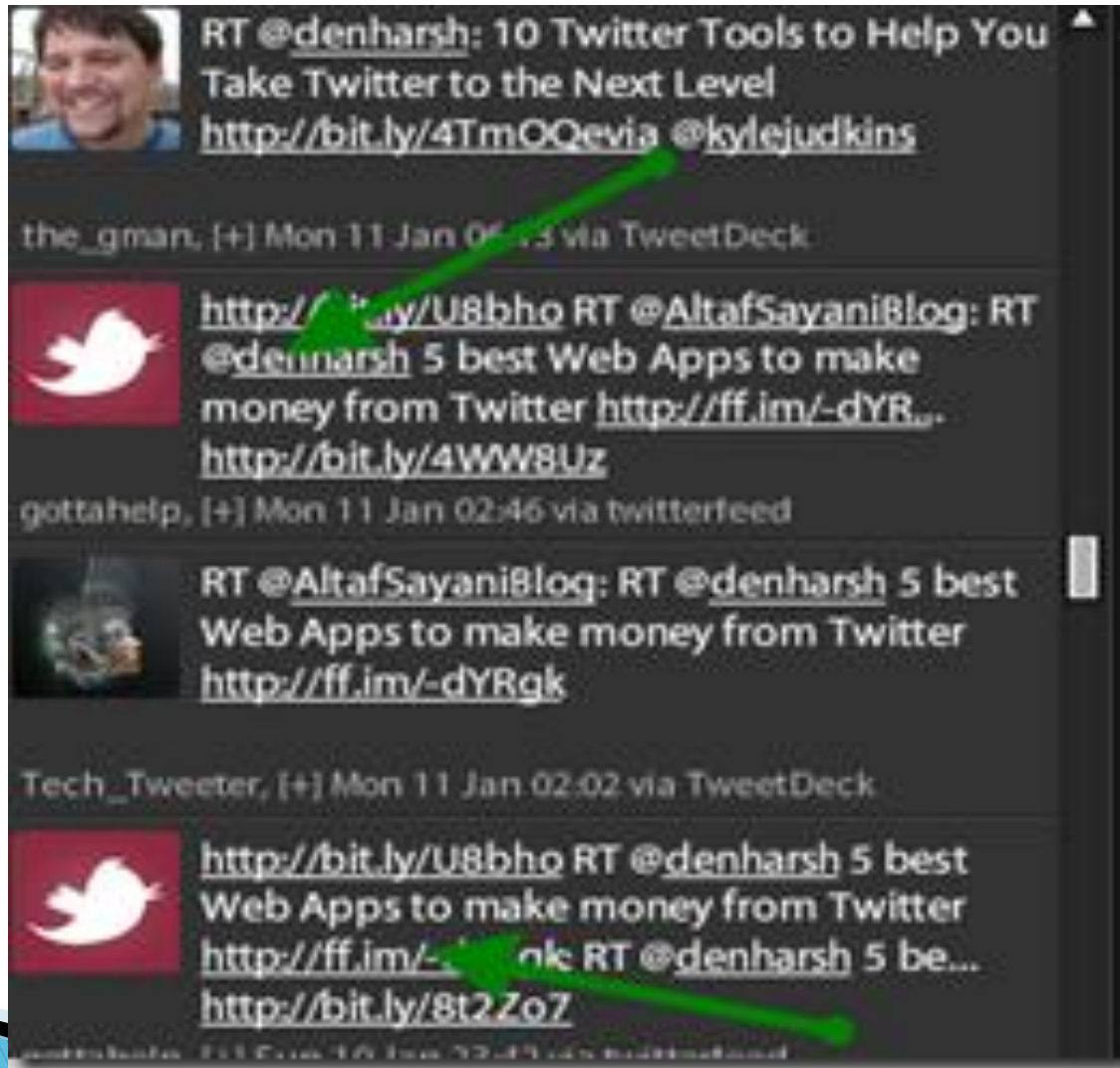


- ▶ All the detected MySpace bots belonged to this category

Bragger

- ▶ Bots that post messages to their own feed
- ▶ Messages vary according to the networks
 - Facebook: status updates
 - Twitter: tweets
- ▶ Spam message is distributed and shown on all the victims' feeds
- ▶ Therefore, the spam campaign reaches only victims who are directly connected with the spam bot
- ▶ 163 bots on Facebook belonged to this category, as well as 341 bots on Twitter

Bragger Example



Poster

- ▶ Bots that send a direct message to each victim
- ▶ This can be achieved in different ways
 - On Facebook, for example, the message might be a post on a victim's wall
- ▶ Most effective way of spamming, because it reaches a greater number of users compared to the previous two.
- ▶ Eight bots from this category have been detected, all of them on the Facebook network.

Whisperer

- ▶ Bots that send private messages to their victims
- ▶ As for “poster” bots, these messages have to be addressed to a specific user
 - Difference: Victim is the only one seeing the spam message.
- ▶ Fairly common on Twitter, where spam bots send direct messages to their victim.



Spam Bot Analysis

	Facebook	MySpace	Twitter
Displayer	2	8	0
Bragger	163	0	341
Poster	8	0	0
Whisperer	0	0	20

Spam Bot Analysis

- ▶ Two kinds of bot behavior were identified
 - Greedy : Include a spam content in every message they send (416)
 - Stealthy: Send messages that look legitimate, and only once in a while inject a malicious message (98)
- ▶ Most spam profiles under observation, both on Facebook and Twitter, sent less than 20 messages during their life span.

Spam Profile Detection

- » Used machine learning techniques to classify spammers and legitimate users.
- » Developed six features to detect whether a given profile belongs to a spammer or not.

Spam Profile Detection

- ▶ FF ratio (R)
 - The feature compares the number of **friend requests** that a user sent to the number of **friends** she has
 - Unfortunately, the number of friend requests sent is not public on Facebook and on MySpace
 - On Twitter, the number of users a profile started to follow is public.
 - $R = \text{following} / \text{followers}$

Spam Profile Detection

- ▶ URL ratio (U)
 - The feature to detect a bot is the presence of URLs in the logged messages
 - $U = (\text{messages containing urls} / \text{total messages})$
- ▶ Only count URLs pointing to a third party site when computing this feature

Spam Profile Detection

- ▶ Message Similarity (S)

-

$$S = \frac{\sum_{p \in P} c(p)}{l_a l_p}$$

- ▶ Friend Choice (F)

-

$$F = \frac{T_n}{D_n}$$

Spam Profile Detection

- ▶ Messages Sent (**M**)
- ▶ Friend Number (**FN**)
- ▶ The Weka framework with a Random Forest algorithm was used for the classifier.

Spam Detection on

► Challenges

- Obtain a suitable amount of data to analyze
- Most profiles are private, and only their friends can see their walls
- Geographic networks discontinued in October 2009
- Used data from various geographic networks, crawled between April 28 and July 8 2009, to test our approach.

► Modifications

- R feature not applicable

Detection on

- ▶ Trained classifier using 1,000 profiles
 - 173 spam bots that contacted honey-profiles
 - 827 manually checked profiles
- ▶ From 790,951 profiles
 - Detected: 130
 - False positive: 7
- ▶ From 100 profiles
 - False negative: 0
- ▶ Low number may be due to the fact that spam bots typically do not join geographic networks

Detection on

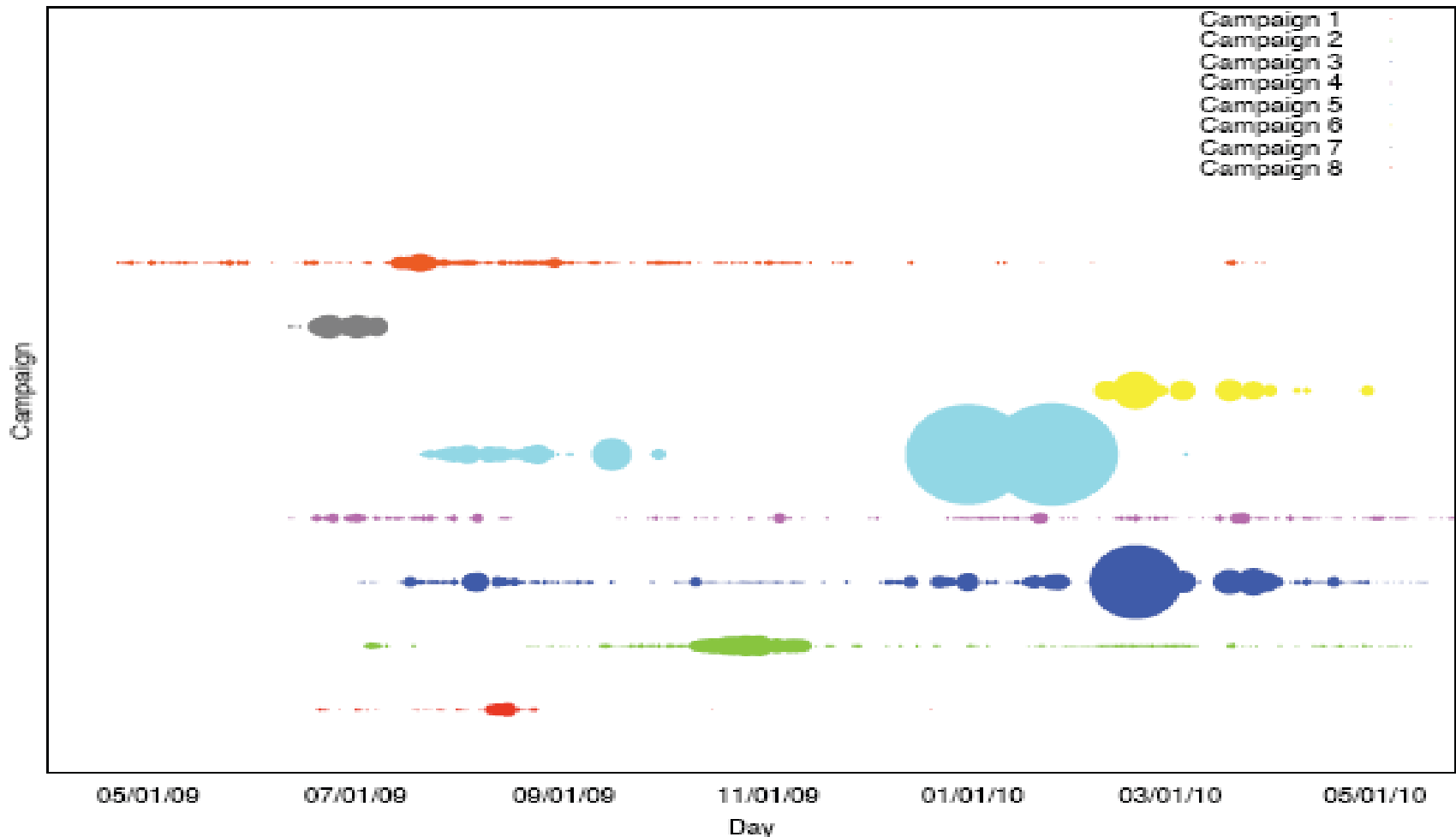
- ▶ Easier to obtain data than Facebook
 - Most profiles are public
- ▶ 500 spam profiles picked
- ▶ 500 legitimate profiles also picked
- ▶ Modifications:
 - Legitimate profiles with a fairly high number of followers (e.g., 300), but following thousands of other profiles, have a high value of R.
 - Therefore, a new feature R'
 - $(R \text{ value}) / (\text{the number of followers a profile has})$

Detection on

- ▶ The main problem faced while building the system was the crawling speed
 - Twitter limited the machine to execute only 20,000 API calls per hour
- ▶ From March 06, 2010 to June 06, 2010, we crawled **135,834** profiles, detecting **15,932** of those as spammers.
 - False positive: 75

Identification of Spam Campaigns

- ▶ Two bots posting messages with URLs pointing to the same site are considered part of the same campaign



Spam Campaigns– Results

#	SN	Bots	# Mes.	Mes./day	Avg. vic.	Avg. lif.	G _c	Site adv.
1	T	485	1,020	0.79	52	25	0.28	Adult Dating
2	T	282	9,343	0.08	94	135	0.60	Ad Network
3	T,F	2,430	28,607	0.32	36	52	0.42	Adult Dating
4	T	137	3,213	0.15	87	120	0.56	Making Money
5	T,F	5,530	83,550	1.88	18	8	0.16	Adult Site
6	T,F	687	7,298	1.67	23	10	0.18	Adult Dating
7	T	860	4,929	0.05	112	198	0.88	Making Money
8	T	103	5,448	0.4	43	33	0.37	Ad Network

A relationship exists between the lifetime of bots and the number of victims targeted

Spam Campaigns– Results

4,7 use a Stealthy Approach

#	SN	Bots	# Mes.	Mes./day	Avg. vic.	Avg. lif.	G _c	Site adv.
1	T	485	1,020	0.79	52	25	0.28	Adult Dating
2	T	282	9,343	0.08	94	135	0.60	Ad Network
3	T,F	2,430	28,607	0.32	36	52	0.42	Adult Dating
4	T	137	3,213	0.15	87	120	0.56	Making Money
5	T,F	5,530	83,550	1.88	18	8	0.16	Adult Site
6	T,F	687	7,298	1.67	23	10	0.18	Adult Dating
7	T	860	4,929	0.05	112	198	0.88	Making Money
8	T	103	5,448	0.4	43	33	0.37	Ad Network

A relationship exists between the lifetime of bots and the number of victims targeted

Success of a Campaign

- ▶ $G_c \geq 0.5$: high success probability

$$G_c = \frac{M_d^{-1} \cdot S_d}{((\sqrt{M_d^{-1} \cdot S_d}) + 1)^2}, \quad 0 \leq G_c \leq 1.$$

Md: avg # of sent Mes/day

Sd: Ratio of actual spam msg

Conclusion

- ▶ Social networking sites attracts spammers
 - ease of reaching these users
 - possibility to take advantage of the information stored in their profiles
- ▶ Created a population of 900 honey-profiles on three major social networks and observed the traffic they received
- ▶ Developed techniques to identify single spam bots, as well as large-scale campaigns.
- ▶ These techniques can help social networks to improve their security and detect malicious users.

Our Proposals

Reverse Social Engineering Attacks

twitter

facebook

- ▶ Attacker does not initiate contact with victim
- ▶ Victim is tricked into contacting the attacker
- ▶ Techniques of abuse of friend-finding features

Find More Friends

Your friends use the friend finder. Have you tried it?

 **Yossi** found 34 friends

 **Shoshana** found 5 friends

 **Shlomi** found 1 friend

 Your Email

Email Password

Show Contacts

friendfeed

Home

- ▶ A List
- ▶ Chicas
- ▶ IM
- ▶ LA
- ▶ Mashable
- ▶ Misc
- ▶ Personal
- ▶ Professional
- ▶ San Diego

Me

Rooms

- The Apple Room
- 1 minute ago

FriendFeed Feedb...

- 3 minutes ago

Apps

- 11 minutes ago
- prefs | 26 more »

Everyone

Friends

-

Find your Twitter friends on FriendFeed

See who's already using FriendFeed and easily subscribe to them.

☒ Select/Deselect all

☒

 **Tim Street** @1timstreet

☒

 **41235** @41235 (private feed)


☒

 **Brenda Young** @4byoung

☒

 **//de** @9keme

☒

 **Atherton Bartelby** @abartelby

☒

 **Abbas Haider Ali** @abbashaiderali

Reverse Social Engineering Attacks

- ▶ Once a reverse social engineering attack is successful, a wide range of attacks such as persuading victims to click on malicious links, blackmailing, identity theft, and phishing can be done!
- ▶ Can bypass current behavioral and filter-based detection techniques that aim to prevent wide-spread unsolicited contact.
- ▶ Third, if the victim contacts the attacker, less suspicion is raised, and there is a higher probability that a social engineering attack (e.g., phishing, a financial scam, information theft, etc.) will be successful.

References

- ▶ G. Stringhini, C. Kruegel, G. Vigna, “Detecting Spammers on Social Networks”, Proceedings of ACM ACSAS’10, Dec, 2010.
- ▶ A. H. Wang, “Don’t Follow me: Twitter Spam Detection”, Proceedings of 5th International Conference on Security and Cryptography, July, 2010.
- ▶ K. Chellapilla and A. Maykov, “A taxonomy of JavaScript redirection spam,” in Proceedings of the 3rd International Workshop on Adversarial Information Retrieval on the Web, 2007.

