Detection of Spam, Unwelcomed Postings, and Commercial Abuses in Social Networks

Vincent Granville, Ph.D.

Chief Architect, Executive Director, Founder

www.analyticbridge.com

vincentg@datashaping.com

November 10, 2011

Types of Spam

- To promote scams, distribute fraud (viruses) or advertise illegal products (online pharmacy)
  - Broken English, use of redirects
  - Easy to detect via email address of member, on sign-up
- Commercial spam
  - Can be relevant to your network, or borderline
  - Uses words such as ‘best’, ‘free’, and catchy titles
  - Posted by interns in India, sometimes by corporate employees
  - Examples on AB: to promote SEO services, or resume writing services, or even commercial software
  - Associated with cross-posting and repetitive posting
  - Sometimes the poster seems to be a very attractive young person

Scammer Profiles (1/3)

- Mismatch between email address and name
  - Text mining to detect mismatch (distance between name and email address)
  - Example: name = Rosalie, email = AmandaLove18@hotmail.com
- Always use free email accounts
- Country is PH, SN, etc. or no country provided,
- Sometimes US (false negatives)
- Also, these members try to get many friends very fast

Scammer Profiles (2/3)

- Profile red flags
  - Select all options in all multiple-choice questions (I am a … student AND retired AND professor AND executive AND manager AND …)
  - Or provide no info except to mandatory questions
  - Repetitive text or answers (‘I am a nice girl’), repetitive links (outside a white list of domain names), abused redirect URLs (bit.ly)
  - Unusual sentences (e.g. ‘will tell you later’; create list of these KWs)
  - Gibberish, stuff like ‘frdsed’ (check these 6 keys on your keyboard)
- Caveats
  - Chinese people tend to have profiles triggering spam flags; need to be treated separately
  - Some of the best members also have very discreet profiles – do a Google search on their name / email (see next slide)

Scammer Profiles (3/3)

- Google the email address of the suspect new member
  - Web crawling / text mining: analyze search results
  - Count occurrences of email address in search results (either 0 or high)
  - Count occurrences where email address associated with spam (Viagra, casino, etc. – use keyword black list for detection)
  - Email address has many digits (OK if digits represent year or zip-code)
  - Sometimes associated with cross-posting
- Naïve Bayes to score email addresses
- Most cases can be automatically detected on sign-up
  - In real time or close to real-time
  - Before they start being active (time to first post is usually short)

Scammers: Solutions (1/2)

- Quarantine new members (no posting for 48 hours)
- Approve each post (not scalable?)
- Real-time sign-up approval algorithm
- Ask detailed questions on sign-up form
  - Mix of mandatory and non-mandatory questions
  - Captcha (to prevent bots from signing up)
  - KDNuggets idea
- Request email address confirmation
  - Drawback: request for email confirmation can go to junk mail

Scammers: Solutions (2/2)

- Allow rejected subscribers to have their case reviewed
  - Reduce false positives
  - Scammers are unlikely to follow up
  - But don't tell a new member she is rejected, and don't tell why
- Create decoy accounts (honey pots) to catch spam
- Allow members to report spam
- Show featured content on top
  - Non-detected spam gets buried underneath
- Disallow features such "post to the entire group"
  - Only promoted members can use this feature

Commercial Spam (1/2)

- Ranges from almost irrelevant to extremely relevant
- Performed by corporate employees, untrained “paid-to-post” interns, or by small businesses
- Delicate situation
  - Who wants to fight against IBM?
  - Content is not irrelevant, poster might have many connections, be respected and create viral marketing for you as well.
  - Viral marketing occurs via tweets or updates “John M from XYZ (big company) posted on Vincent’s group … Excel spreadsheets are full of errors…”. If John has 2,000 connections, it will drive traffic to your network.
  - Solution: don’t feature the post if it’s a free sales pitch

Commercial Spam (2/2)

- Easy to detect
  - Significant cross-posting across multiple networks or LinkedIn groups
  - Cross-posted **on your network** (e.g. LinkedIn group) by multiple members, and by members with a large number of connections, or paid members in India with very few connections [limit number of posts to 5 per day per member?]
  - Very different from a great article that 500 people “like” on Facebook
  - Typically announcing a free webinar or white paper
  - Or (if a small business): SEO services, resume services, tutoring (“we will write your PhD”), free software, etc.
  - Need to identify categories of services with high rate of spam.

Identifying Content Duplication

- Spammers massively cross-post the same content
  - Need to identify duplicated content to help fight spam
  - Easy if exactly identical content cross-posted by same member, more difficult if multiple members involved, or slight changes in content
- Text Mining Algorithm
  1. Stem each document using a **stemming algorithm**, remove **stop words**, (using stop word list), fix typos, alphabetically order tokens
  2. For each stemmed document, identify list of top rarest tokens, and store full stemmed document + URL + rarest tokens in centralized DB
  3. Rarest tokens (or pairs of rarest tokens) used as key in the database
  4. When you check a document, search for potential duplicates using rarest tokens, then compare documents sharing many rare tokens

Don’t Become the Spammer!

- Categorize your members (via sign-up questions)
  - Create a member taxonomy
- Send targeted e-mail blasts to each member category, rather than generic e-mail blasts to everybody
  - Recruiters and non-recruiters receive different messages
  - Consider offering geo-targeting
- Carefully select your advertisers
- Allow members to receive weekly digests rather than 100 messages per month
- Be very careful about selling or swapping mailing lists

Special Issues (1/2)

- Dormant accounts
  - Potential time bombs? (Botnets)
  - Delete profiles after 6 months of inactivity (warn user before deletion)
- Web log spam
  - Manipulation of web traffic stats to increase page view counts, Facebook “likes” and other popularity metrics, using web robots to generate page views, or “paid-to-like” people in India
  - Purpose: getting a post to artificially show up with a top position in the list of most popular postings, to get more traffic
  - Solution: do not make page-view count data publicly available, for each post show number of unique visitors rather than page views (more difficult to cheat)

Special Issues (2/2)

- Fake profiles
  - Rings (clusters) of 5 members belonging to a same (fake) company, recommending each other – could be just one person
  - Spammers copying a well respected profile to avoid detection
- Link posting
  - Do not allow postings that are essentially one link from outside a white list of domain names
- Spammers replacing URLs with Google keywords
  - Example of post: “search for ‘wealthy singles America’ on Google”
  - The spammer posting this message managed to get a #1 position for her URL on Google organic (or paid!) for the keyword in question
  - Used to avoid detection

Examples of Spammers

sastrainingonline	sasolinetrngn8@gmail.com
Josh Cooksey	JCooksey@greenkeyllc.com
yaosangjian	zhaoxufang11@gmail.com
Daniel Martin	danielcdmartin@gmail.com
liuliu	ll131363@sina.cn
Jeannie J. Astudillo	daniel.zoleta@gmail.com
Alice Johnson	lescheal@yahoo.com
a2zonlinetrainig	azonlinetraining@gmail.com
Elsa	e.lsasom.0@gmail.com
Lou Lamb	LouLamb333@gmx.com
prisca sammuel	prisca_sammuel@yahoo.com
anne desmond	annedesmond31@rocketmail.com
rose mary	meryk25@yahoo.com
serena li	lisiqin_007@yahoo.com
Rose Grandson	rosegrandson77@yahoo.ca
Juliana	vaye16faye@yahoo.com
lovejuliana	vaye17faye@yahoo.com
Oxymoron	onlinebookkeepingsoftware@ymail.com
Rajesh Kannan	rajesh.submission@gmail.com
Branden Berardi	conn.elly.r0447@gmail.com
kdsfjklds	kdsfjklds@126.com
Paul Jacob	topjacob@gmail.com
peijz	peijz@yahoo.cn
juliajhon	vebnest12@gmail.com
Zachary01	andrewhudson101@gmail.com
Betty J. Brewster	amalindikasampath@gmail.com
vera	tony20_you@yahoo.com
Larry P	larryp7639@gmail.com
Fred Bergman	denswerter@yahoo.com
Tim Fellert	killensgreg@yahoo.com
Eliza Dushku nude	sunbeames@yahoo.com

Conclusion

- Two types of social networks:
  - **Spam infested**: If you don't fight spam using appropriate methodology, your network will eventually contain nothing but spam and spammers, and it will be impossible to fix the problem
  - **Spam eradicated** (similar to small pox): If you fight spam using appropriate methodology from the beginning, eventually no spammer will ever waste her time on your network, your network will become spam-free, and you won't even have to bother about spam detection anymore