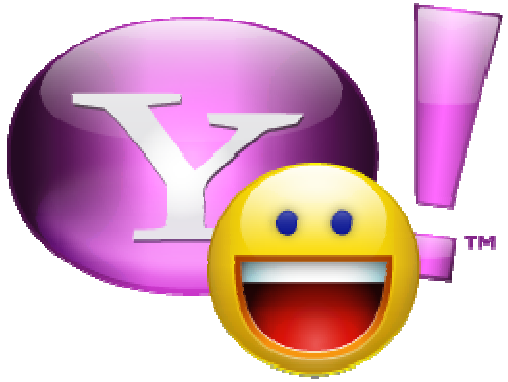
A Fractal Approach to Social Networks Spam Detection

Catalin & Carmen COSOI

Intro



- Social Networks
- Scale Free Networks
- Make some connections
- Draw some conclusions

Twitter

- Huge growth in the last years
- Approximately 7 mil tweets / hour
- In Sept 2008, Twitter had a growth of 343%
  - (sept 2007 vs sept 2008. according to Nielsen-Online)

In Feb 2009, the growth was of 1382% (feb 2008 – feb 2009, according to Nielsen-Online)

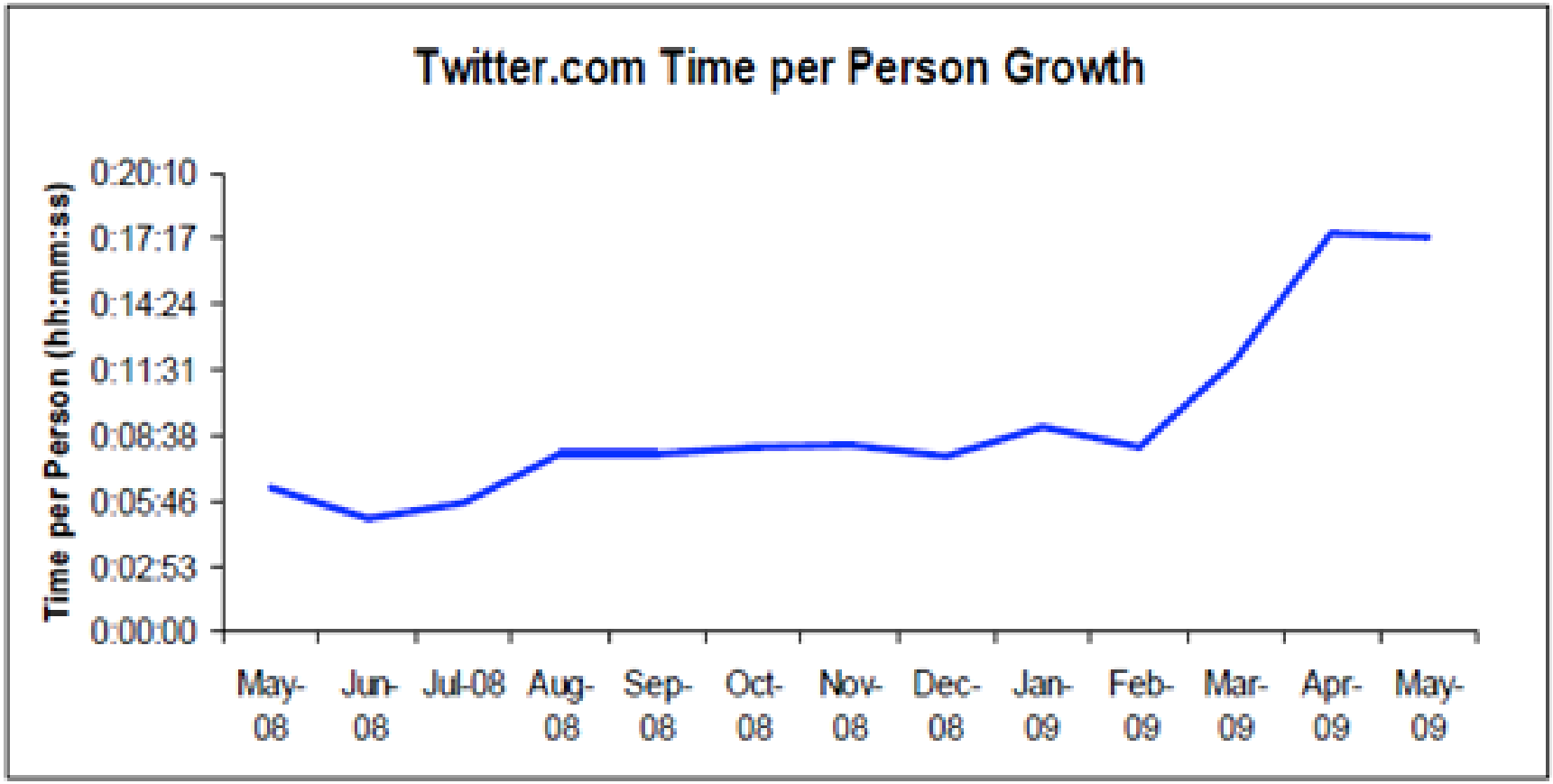
Age Group	Unique Audience	Composition %
2-17	250,000	3.6
18 – 24	**	**
25 – 34	1,379,000	19.6
35 – 49	2,935,000	41.7
55+	1,165,000	16.6
65+	477,000	6.8

source: Nielsen NetView, 2/09, U.S., Home and Work

**These demographics have insufficient sample sizes

bitdefender

Time on Site

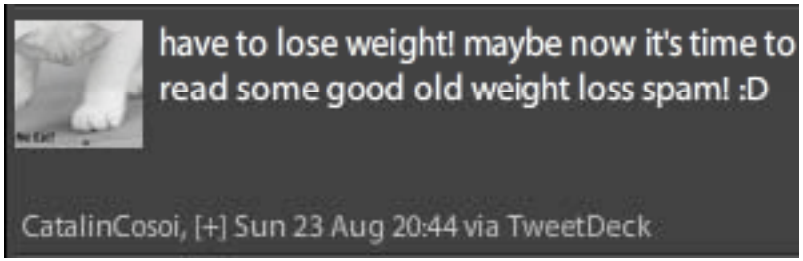


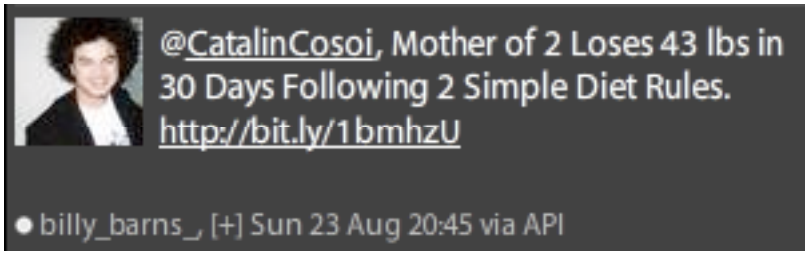
Source: The Nielsen Company

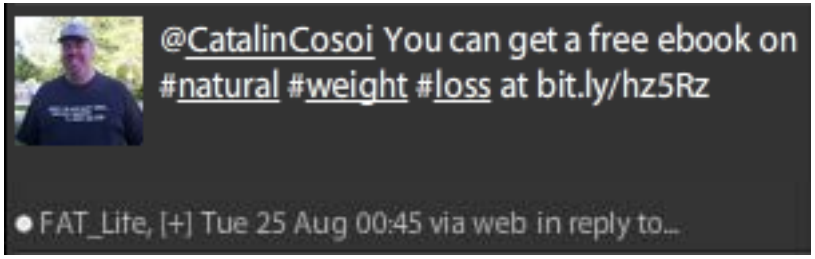
bitdefender

Twitter Spam

- Following Spam
- Tweet Spam
- Direct Message Spam
- Trending Subject Spam
- Reply/Trackback Spam
- ReTweet Spam







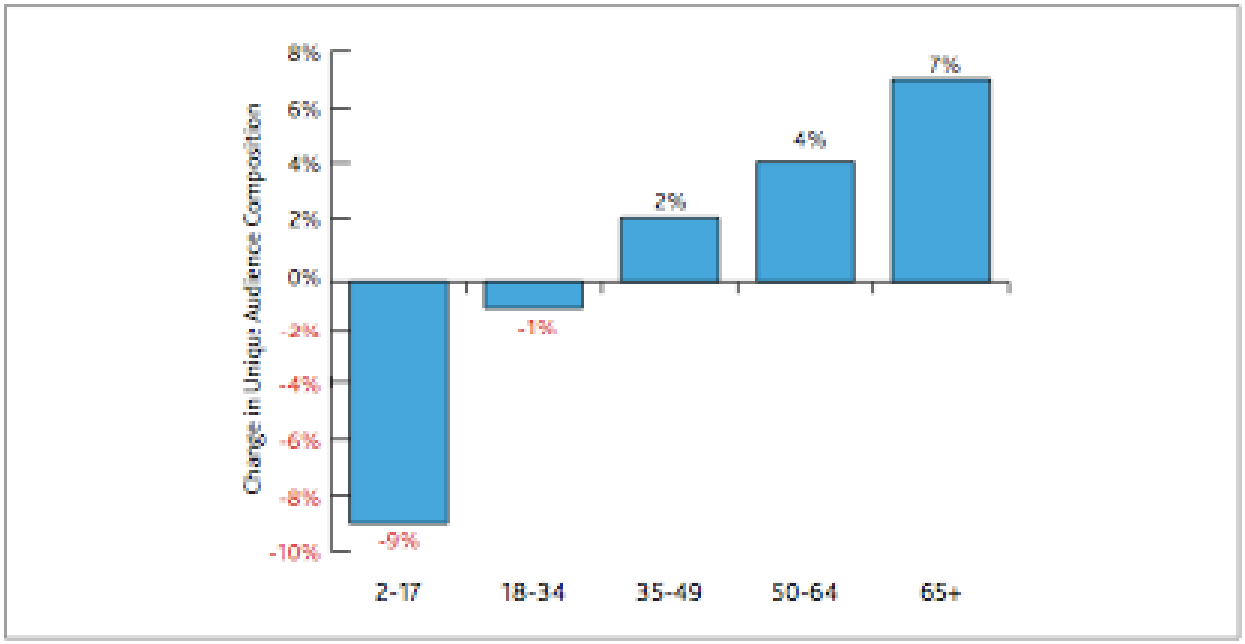
bitdefender

Is it hard to send spam on twitter?

bitdefender

Facebook

- Facebook Sept 2007 to Sept 2008 – 116% growth
- Was ranked in the first place as the social network with the most time spent on

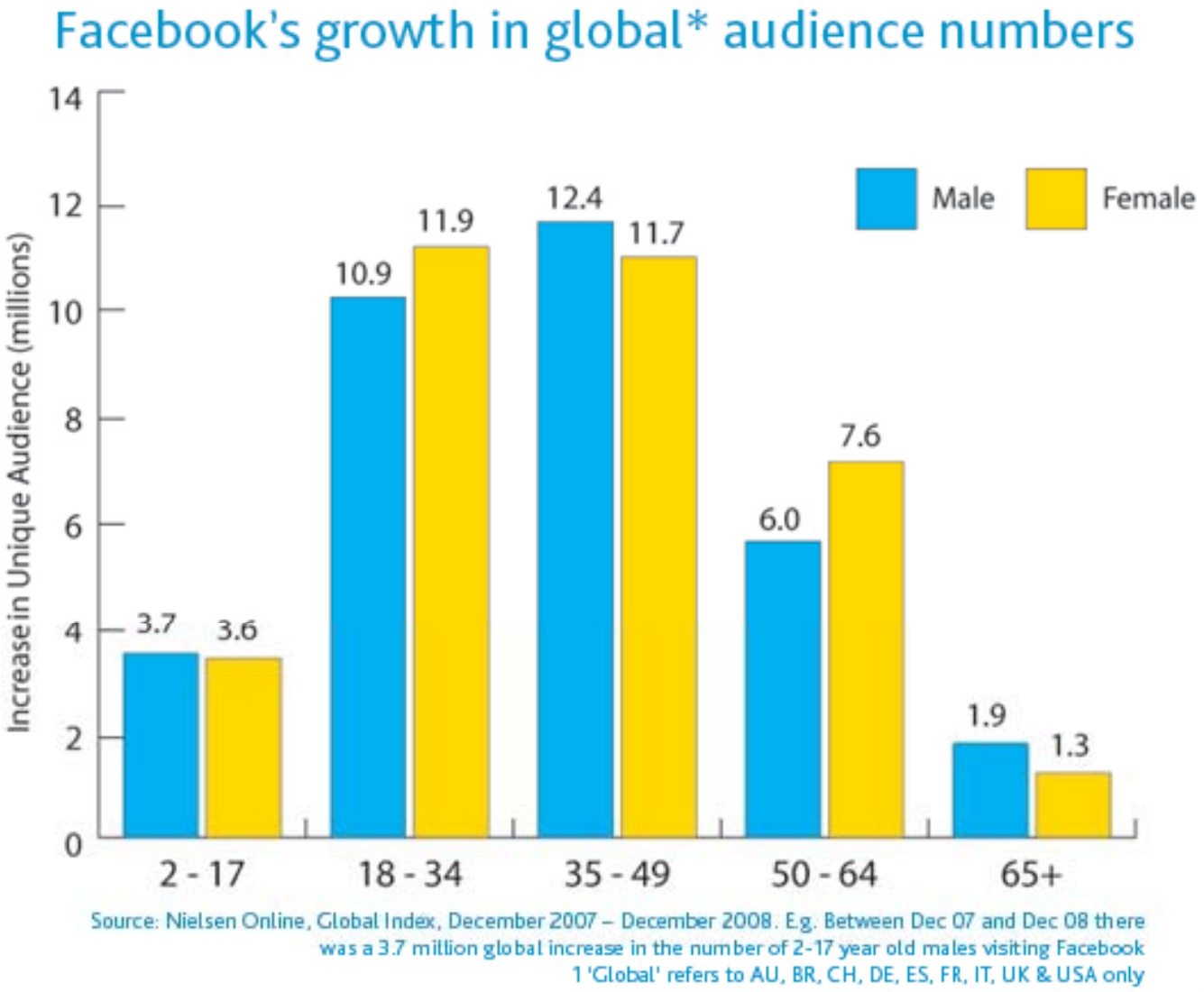


Source: Nielsen Online, Custom Analytics, December 2007 – December 2008. E.g. Between Dec 07 and Dec 08, the share of the online global[1] audience to 'Member Community' sites accounted for by 2-17 year olds decreased relatively by 9%

[1] "Global" refers to AU, BR, CH, DE, ES, FR, IT, UK & USA only

bitdefender

Again, facebook



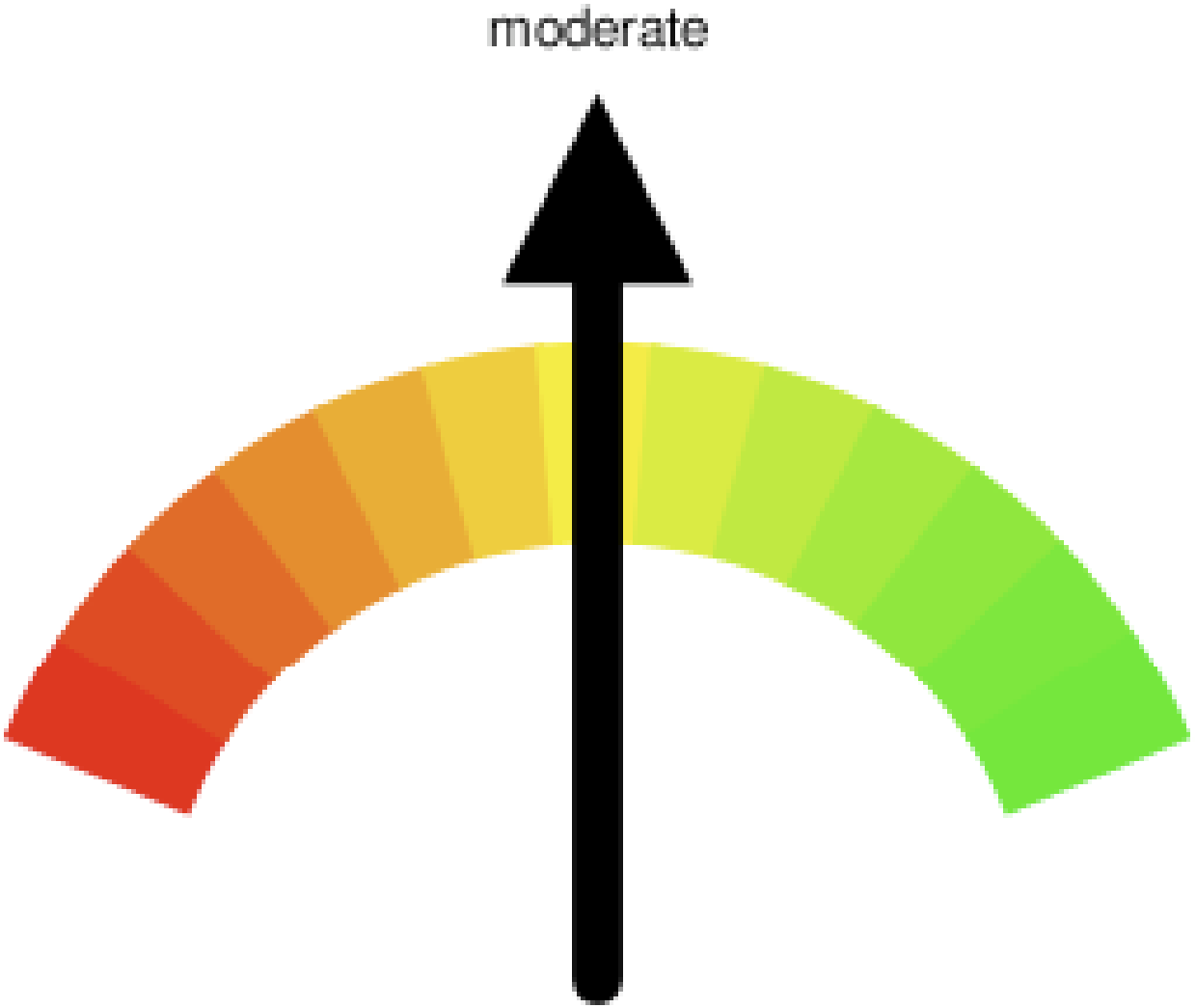
bitdefender

Facebook spam

- Graffiti Spam (they write spam on my wall)
- Anyone in the audience has a porn star friend?
- Status Spam (personal graffiti)
- Comment Spam

bitdefender

Is it hard to send spam on Facebook?



bitdefender

LinkedIn

Average Age	41
Household Income	$109,703
Male	64%

	Comp	Index	Rank
HHI $100K+	53.5%	210	1
Own Smartphone/PDA	34%	430	1
College Grad/Post Grad	80.1%	156	2
Business Decision Maker	49%	185	13
EVP/SVP/VP	6.5%	317	4

http://www.slideshare.net/erickschonfeld/linkedin-demographic-data-jun08-presentation

LinkedIn Spam

- Spam Profiles (high google rankings)
- Spam comments in LinkedIn groups
- Spam “friend” invitations

Purchase Viagra Purchase Viagra no Prescription

Purchase **Viagra** Sildenafil Citrate No Prescription

Canada | Pharmaceuticals

Current	• **Purchase Generic Viagra (Sildenafil 100/50mg) Online** at **Pharmaceutical eConsulting**
Connections	**0** connections
Websites	• Buy Viagra - >> CLICK HERE! << • Purchase cheap Viagra no rx • Order Generic Viagra No Rx
Public Profile	http://www.linkedin.com/pub/purchase-viagra-purchase-viagra-no-prescription/15/7b9/a9a

bitdefender

Blogs

- We all know what blog spam is…
- Is blog spam a real issue?
- Email and blog spam come from the same ips (mostly)
- On my personal blog, spam comments vs legit comments rate per day is: 150 / 0 (division by 0)
  - Spam blogs
  - Comment spam
  - Trackback spam
  - Weird spam

Applied Graph Theory

- Node1 = random (1, N)
- Node2 = random (1, N) | Node2 != Node1
- AddEdge(Node1 , Node2)
- Connections are made randomly

- Pure mathematics, mostly concerned with the combinatorial properties of artificial constructs
- Oriented towards design and engineering

Scale Free Networks

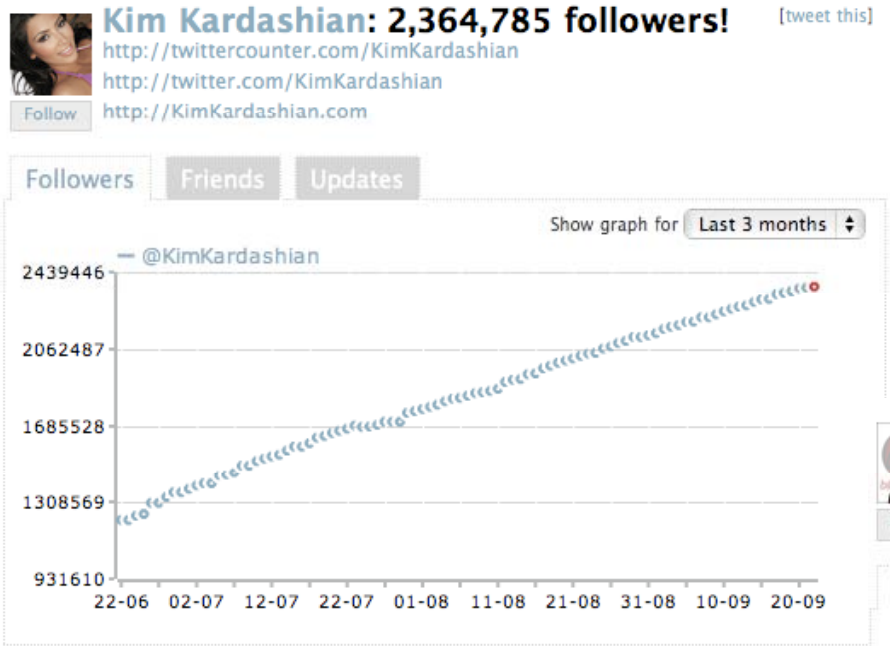
Scale-free network

From Wikipedia, the free encyclopedia

A **scale-free network** is a network whose degree distribution follows a power law, at least asymptotically. That is, the fraction $P(k)$ of nodes in the network having k connections to other nodes goes for large values of k as $P(k) \sim k^{-\gamma}$ where γ is a constant whose value is typically in the range $2<\gamma<3$, although occasionally it may lie outside these bounds.

Scale-free networks are noteworthy because many empirically observed networks appear to be scale-free, including the protein networks, citation networks, and some social networks.[*citation needed*]

Scale Free Networks – Rule #1



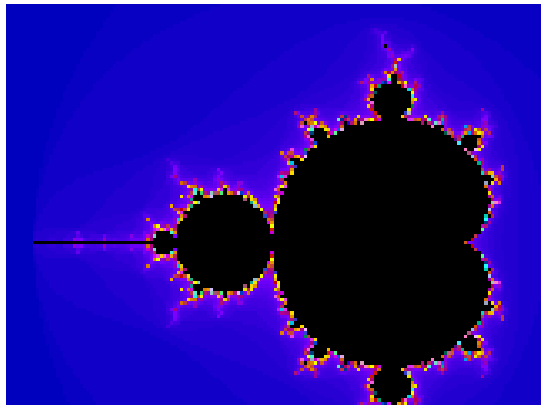
Social Connections are not random. Nodes with many connections will attract new connections easier than nodes with fewer connections

"The rich get richer and the poor get poorer". Proverb provided by wikipedia ☺

bitdefender

Graph Theory vs. Scale Free Networks

- Scale Free Networks accept growth (edges and nodes are added dynamically)
- Scale Free Networks accept preferential attachment
- Tolerant to accidental failures (if we take a few random nodes out, the structure of the network will survive)
- Scale free networks, also called fractal networks manifest the same properties at macro or micro levels.



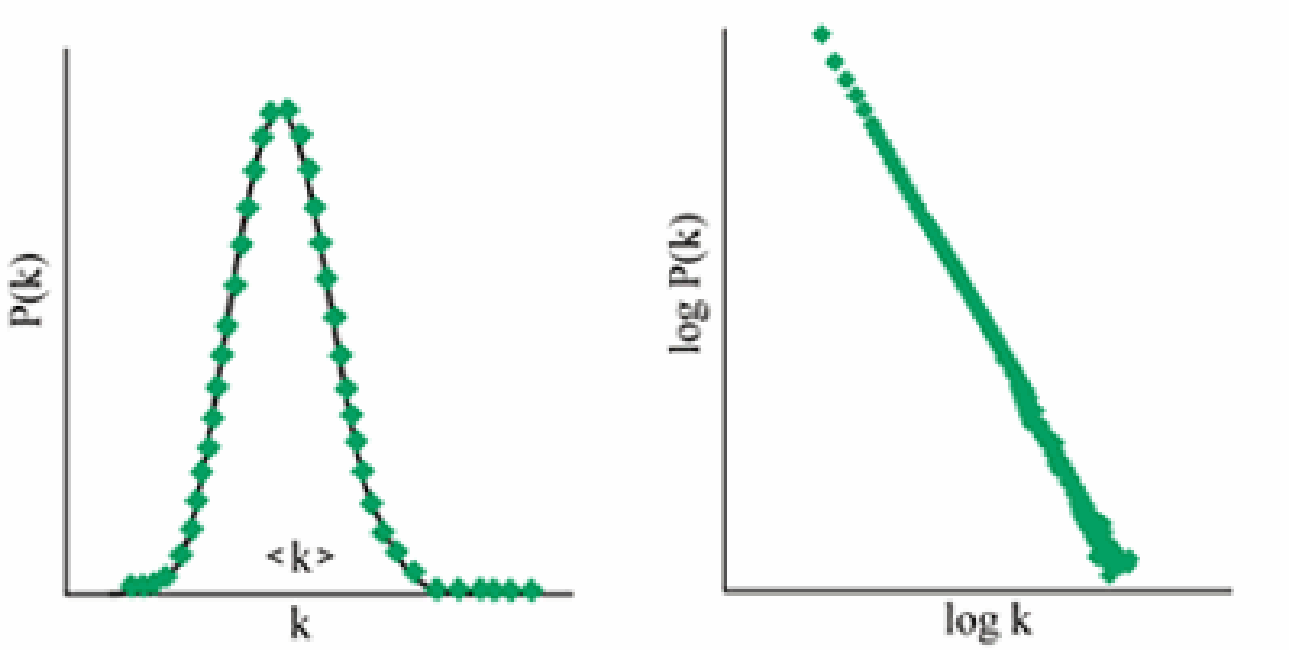
Scale Free Networks - Properties

- Small World
  - Despite their size, there is usually a short path between two nodes
  - “Six degrees of separation” (even less in some networks)
  - … random graphs are small worlds as well
- Clustering
  - Small circles of friends are formed, where every member knows any other member
- Degree Distribution
  - Not all nodes have the same number of edges
  - The probability of node I having k edges can be computed

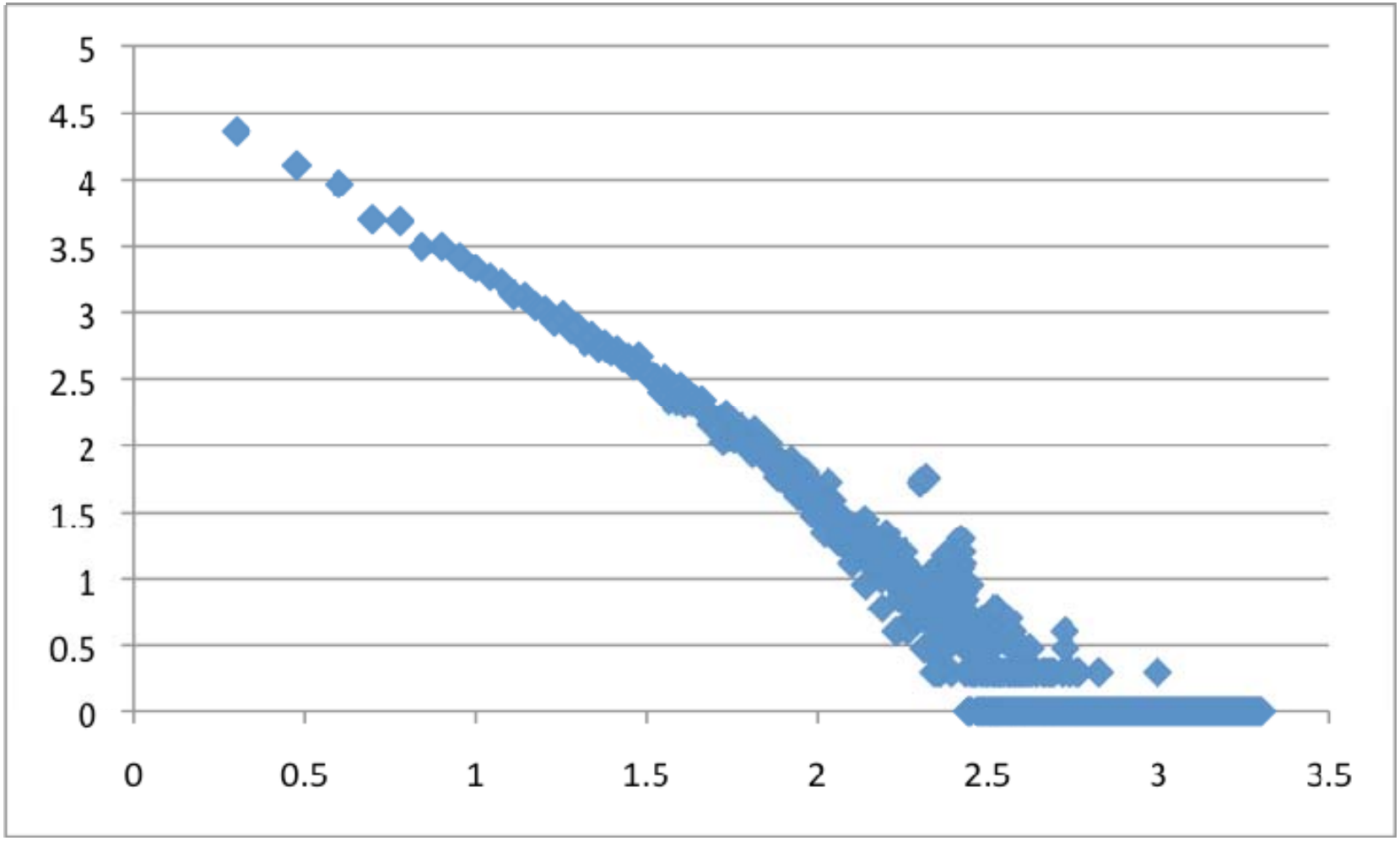
$$P(k_i - linking\_to\_node\_i) = \frac{k_i}{\sum_j k_j}$$

What is the power law?

$$P(k) = ck^{-z}$$

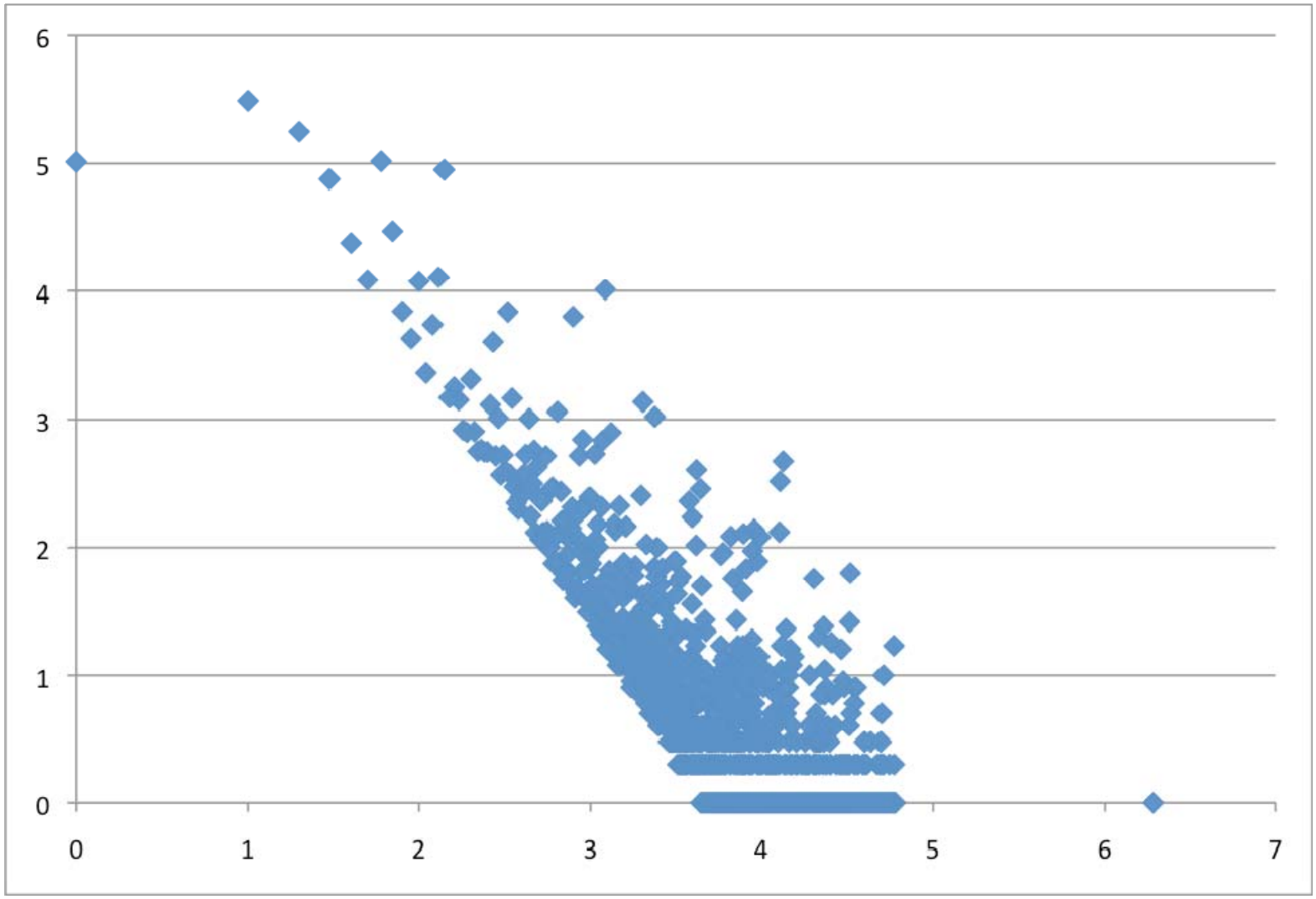


Blogosphere Distribution



bitdefender

Twitter Distribution



bitdefender

Social Networks as Scale Free Networks

- Facebook, Twitter and LinkedIn are scale free networks
- The blogosphere is also a scale free network
- Power Law Distribution: $P(k) = ck^{-z}$
  - exponent for Twitter = 2.71
  - Exponent for Facebook = 2.94
  - Exponent for LinkedIn = 2.31
  - Exponent for the Romanian Blogosphere = 2.65
- Example: aprox 22% twitters have between 0 and 5 friends

Epidemic Profiles

- Random attacks cannot destroy the structure of the network (even if we infect 80%)
- On the other hand, targeted attacks, can!
- Scale free networks are super robust but also super weak!
- 2 Basic models in theoretical epidemiology
  - SIS (susceptive-infected-susceptive)
  - SIR (susceptive-infected-removed)
- 2 basic actions
  - Immunization
  - isolation

Imagine a malware URL from these guys

1	 aplusk Joined 8 months ago http://blahgirls.com Follow	here Tehran	218	3,407	3,660,682
2	 TheEllenShow Joined 1 year ago http://www.ellentv.co... Follow	California Pacific Time (US & Canada)	25	478	3,353,918
3	 britneyspears Joined 1 year ago http://www.britneyspe... Follow	Los Angeles, CA Pacific Time (US & Canada)	433,225	278	3,324,846
4	 cnnbrk Joined 3 years ago http://cnn.com/ Follow	Everywhere Eastern Time (US & Canada)	18	1,123	2,749,368
5	 KimKardashian Joined 6 months ago http://KimKardashian.... Follow	Armenia Pacific Time (US & Canada)	62	1,844	2,393,301

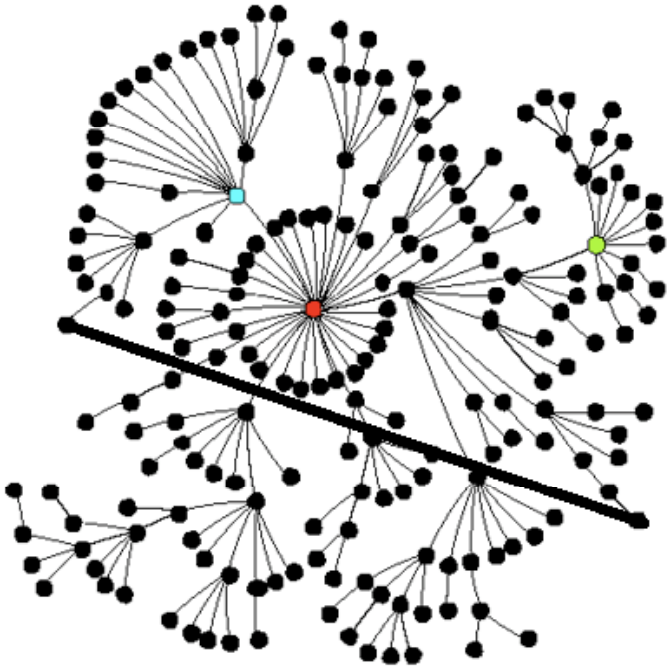
bitdefender

Solutions

- Targeted immunization
- Temporary interrupt communication
- If in social networks the connections are obvious, in blogs are made by “influence”.
- Statistic data + power law distribution can lead to a artificial network fairly close to the real one
- Simulations + Research on the artificial network can provide deep information about the network behavior when attacked

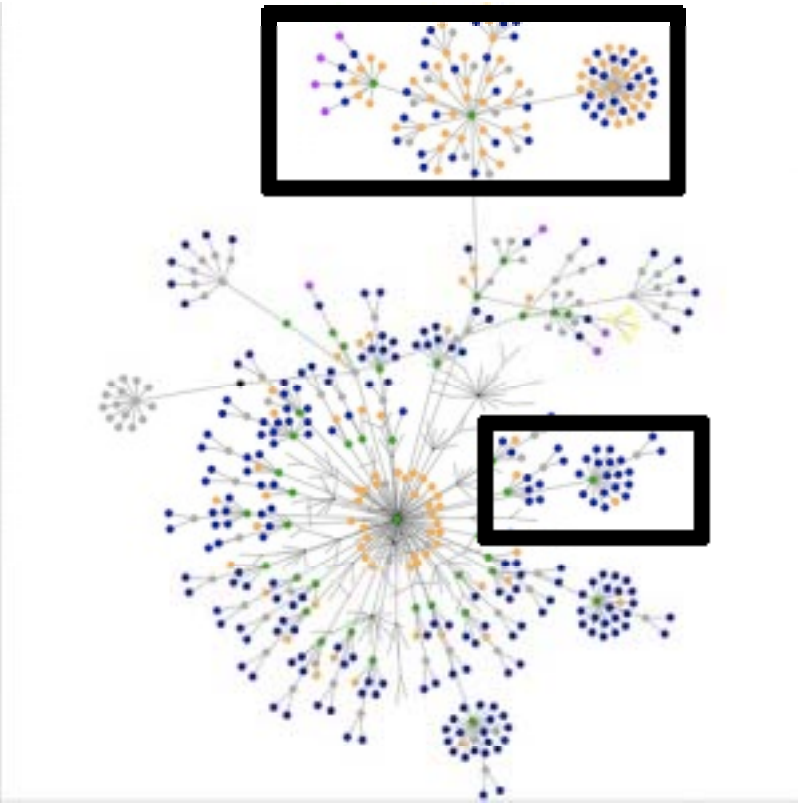
Anomaly Detection

- Improbable connections can be determined
- Improbable small worlds (bot-nets) can be identified
- High connectivity for very young nodes is also improbable
- Since these networks are self similar, they should present the same properties at different sub-component networks



Anomaly Detection

- If two sub-graphs (windows) are too similar or too different, could be another indicator of a anomaly
- Connections are formed and then removed too fast
- Information starts to flow differently



Questions?