Learning Influence Propagation on Personal Blogs

Il-Chul Moon, Dongwoo Kim, Yohan Jo, Alice H. Oh

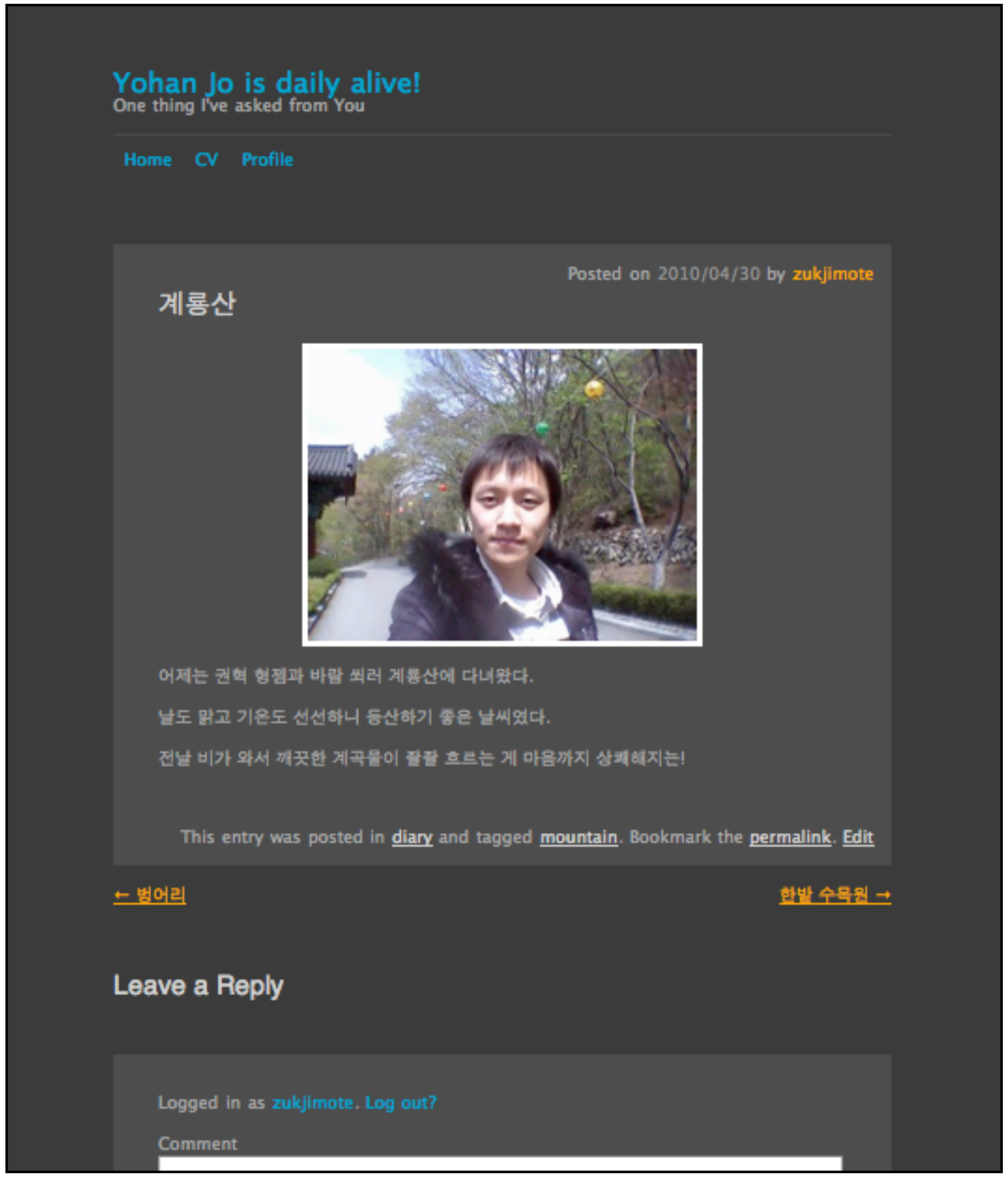
yohan.jo@kaist.ac.kr

KAIST
Daejeon, Korea

In this presentation, I will talk about

- how to measure a blog’s influence to readers
- how to predict the potential influence of a new blog post well

Blog Influence



Yohan Jo is daily alive!
One thing I've asked from You

Home CV Profile

Posted on 2010/04/30 by zukjimote

계룡산

어제는 권혁 형젬과 바람 쐬러 계룡산에 다녀왔다.

날도 맑고 기온도 선선하니 등산하기 좋은 날씨였다.

전날 비가 와서 깨끗한 계곡물이 콸콸 흐르는 게 마음까지 상쾌해지는!

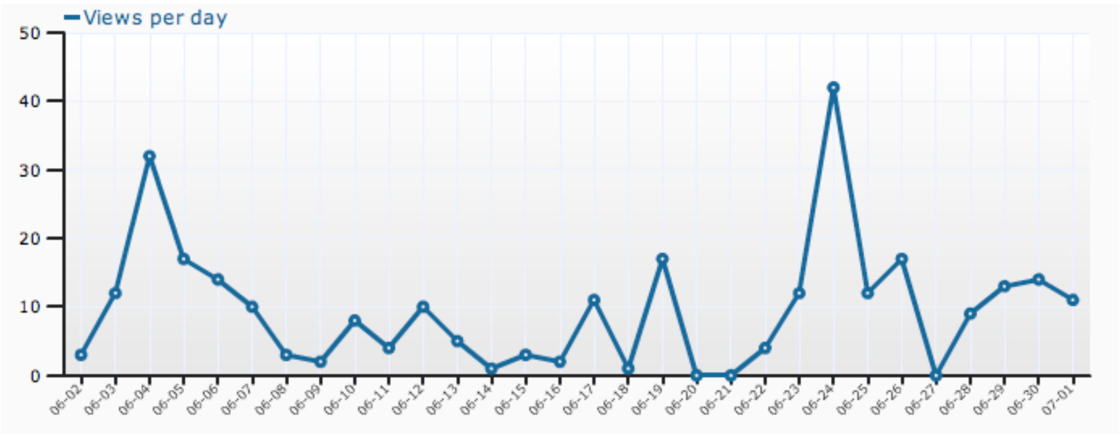
This entry was posted in diary and tagged mountain. Bookmark the permalink. Edit

← 범어리 한밭 수목원 →

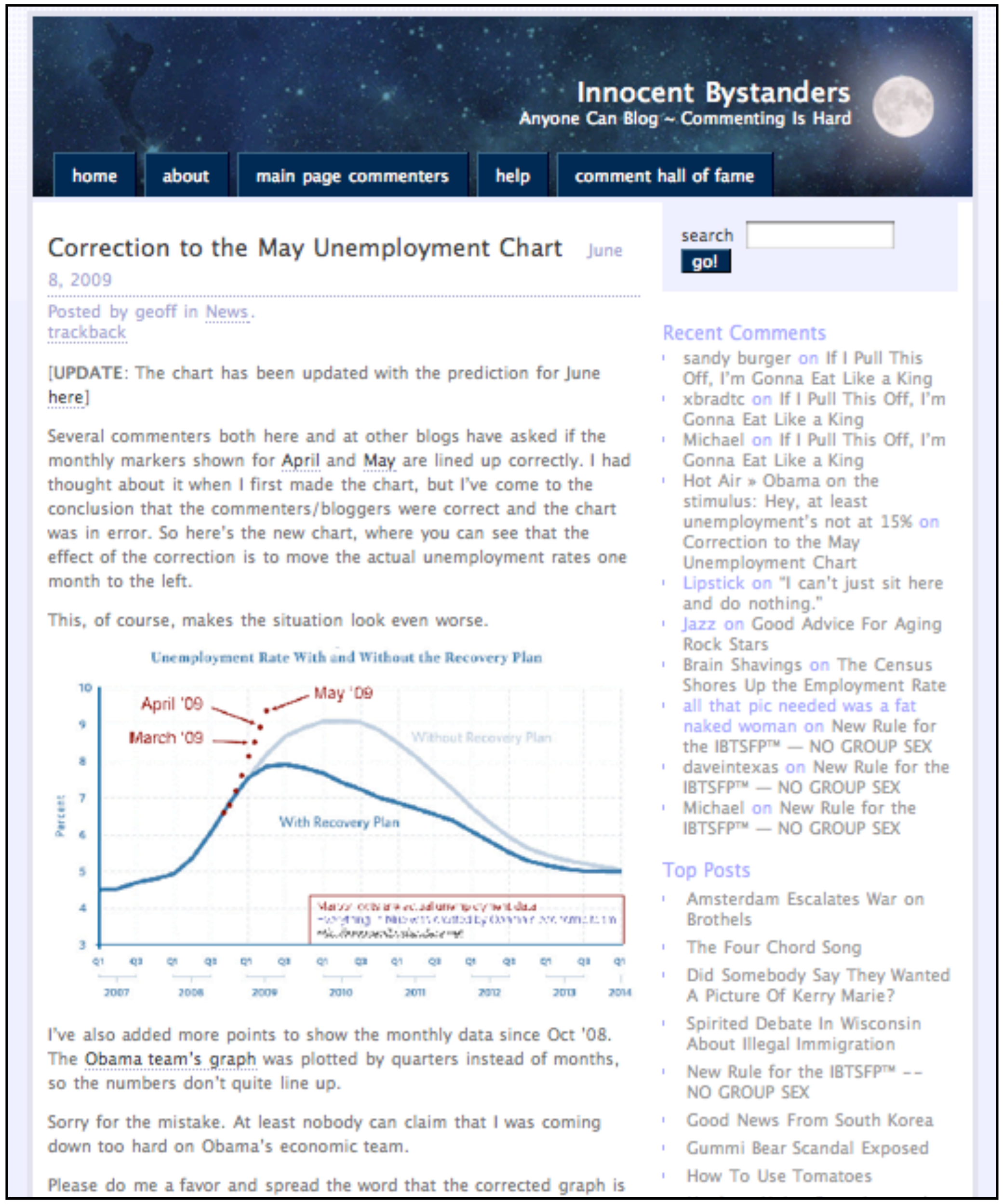
Leave a Reply

Logged in as zukjimote. Log out?

Comment



Blog Influence



Correction to the May Unemployment Chart June 8, 2009

Posted by geoff in News.
trackback

[UPDATE: The chart has been updated with the prediction for June here]

Several commenters both here and at other blogs have asked if the monthly markers shown for April and May are lined up correctly. I had thought about it when I first made the chart, but I've come to the conclusion that the commenters/bloggers were correct and the chart was in error. So here's the new chart, where you can see that the effect of the correction is to move the actual unemployment rates one month to the left.

This, of course, makes the situation look even worse.

I've also added more points to show the monthly data since Oct '08. The Obama team's graph was plotted by quarters instead of months, so the numbers don't quite line up.

Sorry for the mistake. At least nobody can claim that I was coming down too hard on Obama's economic team.

Please do me a favor and spread the word that the corrected graph is

Recent Comments

- sandy burger on If I Pull This Off, I'm Gonna Eat Like a King
- xbradtc on If I Pull This Off, I'm Gonna Eat Like a King
- Michael on If I Pull This Off, I'm Gonna Eat Like a King
- Hot Air » Obama on the stimulus: Hey, at least unemployment's not at 15% on Correction to the May Unemployment Chart
- Lipstick on "I can't just sit here and do nothing."
- Jazz on Good Advice For Aging Rock Stars
- Brain Shavings on The Census Shores Up the Employment Rate
- all that pic needed was a fat naked woman on New Rule for the IBTSFP™ — NO GROUP SEX
- daveintexas on New Rule for the IBTSFP™ — NO GROUP SEX
- Michael on New Rule for the IBTSFP™ — NO GROUP SEX

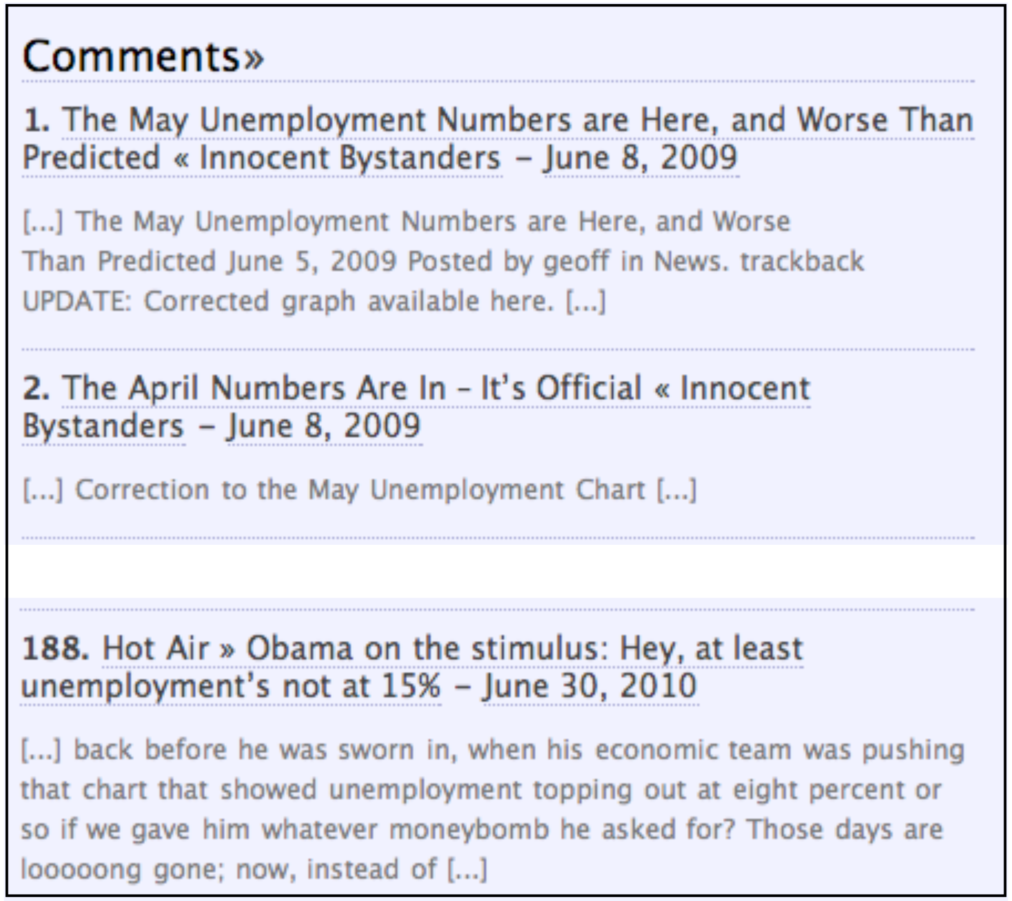
Top Posts

- Amsterdam Escalates War on Brothels
- The Four Chord Song
- Did Somebody Say They Wanted A Picture Of Kerry Marie?
- Spirited Debate In Wisconsin About Illegal Immigration
- New Rule for the IBTSFP™ -- NO GROUP SEX
- Good News From South Korea
- Gummi Bear Scandal Exposed
- How To Use Tomatoes

Hits Measured by WordPress

- 3,783,718 hits since April 1, 2006

2,500 visits / day



Comments»

1. The May Unemployment Numbers are Here, and Worse Than Predicted « Innocent Bystanders – June 8, 2009

[...] The May Unemployment Numbers are Here, and Worse Than Predicted June 5, 2009 Posted by geoff in News. trackback UPDATE: Corrected graph available here. [...]

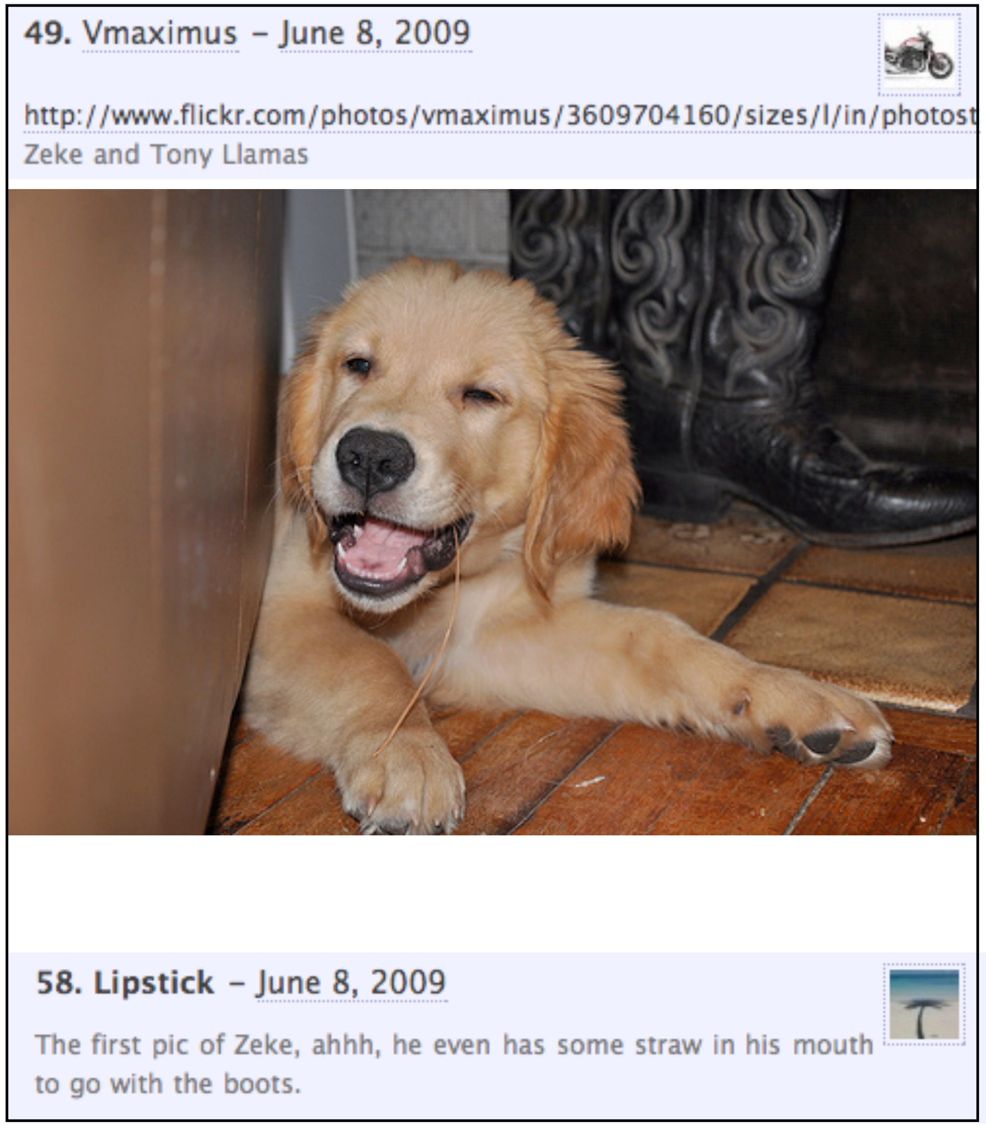
2. The April Numbers Are In – It's Official « Innocent Bystanders – June 8, 2009

[...] Correction to the May Unemployment Chart [...]

188. Hot Air » Obama on the stimulus: Hey, at least unemployment's not at 15% – June 30, 2010

[...] back before he was sworn in, when his economic team was pushing that chart that showed unemployment topping out at eight percent or so if we gave him whatever moneybomb he asked for? Those days are looooong gone; now, instead of [...]

Blog Influence



49. Vmaximus – June 8, 2009

http://www.flickr.com/photos/vmaximus/3609704160/sizes/l/in/photost

Zeke and Tony Llamas

58. Lipstick – June 8, 2009

The first pic of Zeke, ahhh, he even has some straw in his mouth to go with the boots.

Blog Influence

Power of stimulating the readers to express their thoughts in response

The influence is reflected,

- quantitatively, by the **network position**
  e.g. number of people influenced
- qualitatively, in the **content**
  e.g. similarity of the topics

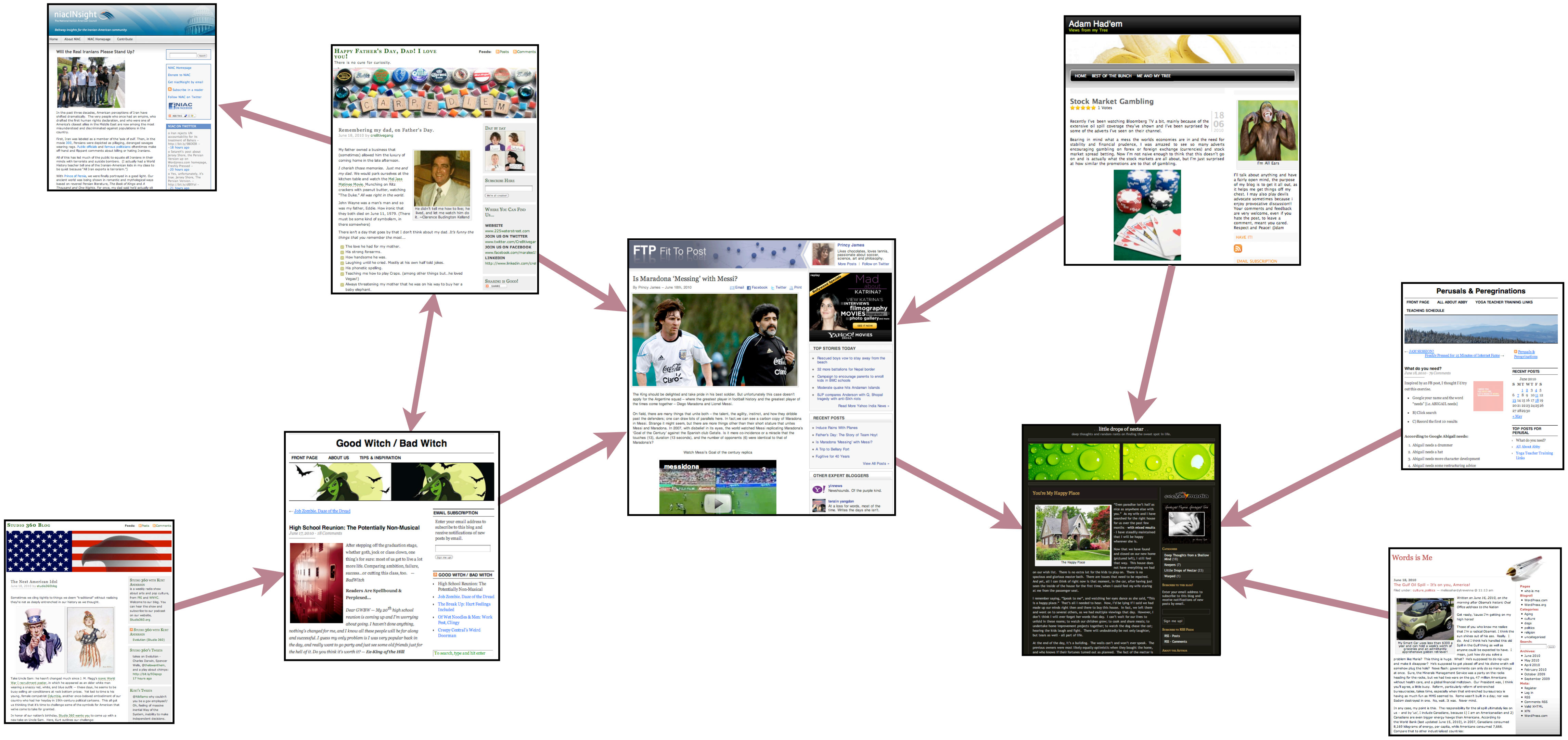
Content Analysis

Apply the author-topic model to the blogs [Rosen-Zvi, 2004]

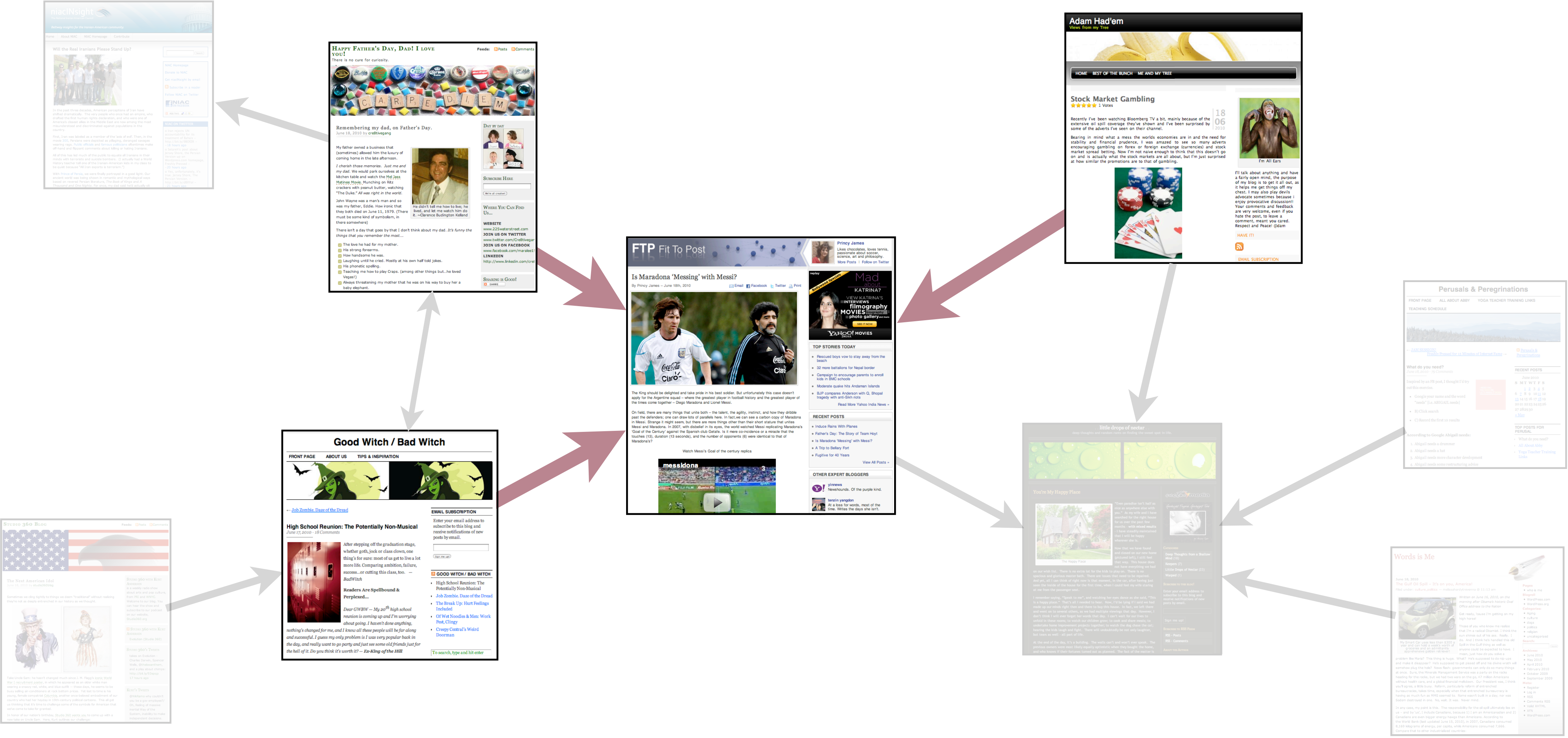
Topic	Highly Related Words
1	gold, wave, market, term, short, cycle
2	money, tax, government, fund, pay, financial
…	
50	school, student, university, education, class

Blog	Topic					
	1	2	3	4	…	50
1	0.011	0.324	0.003	0.008		0.003
2	0.250	0.007	0.012	0.009		0.011
…						
4,165	0.009	0.015	0.003	0.010		0.363

Network Analysis

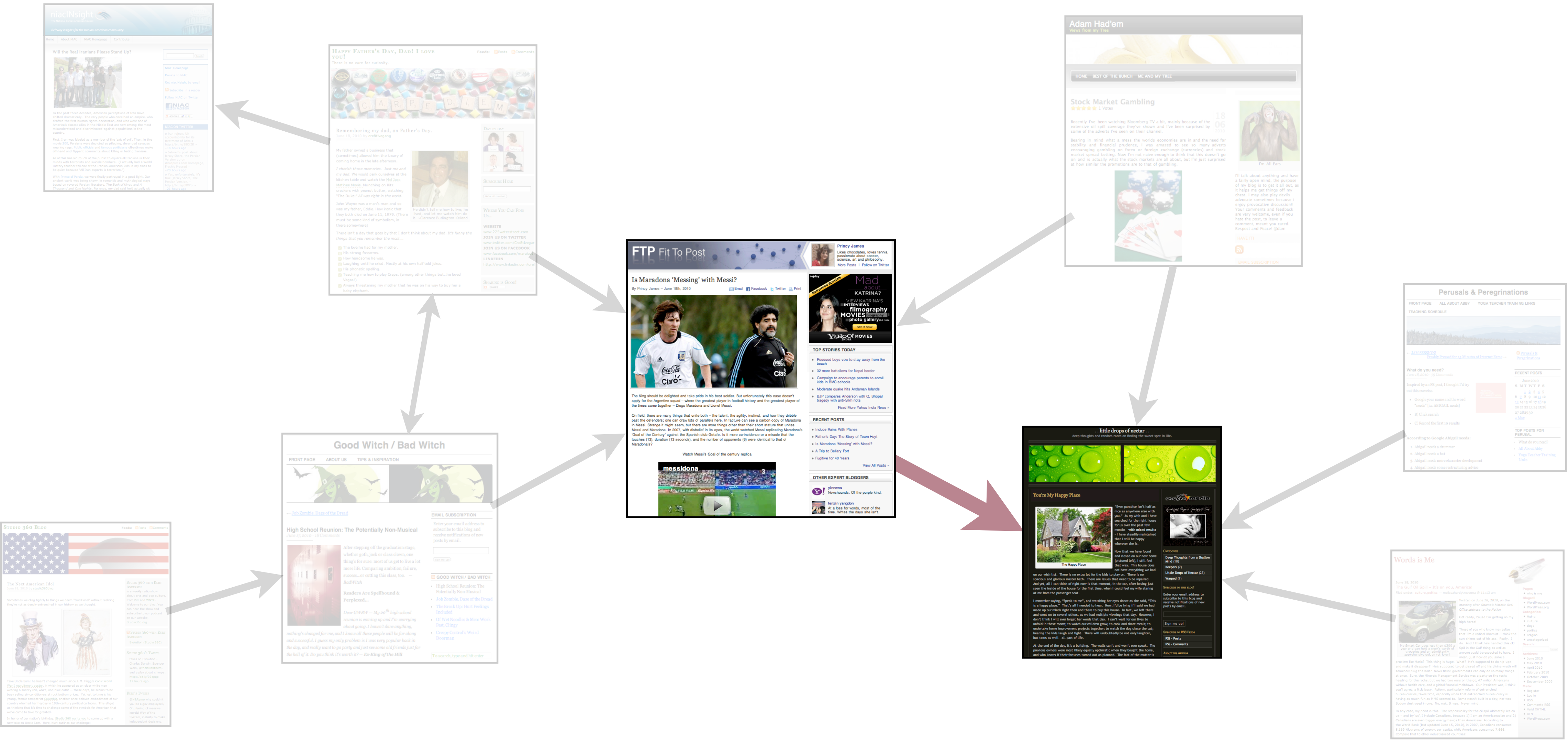


Network Analysis


Happy Father's Day, Dad! I love you!
Adam Had'em
Stock Market Gambling
FTP Fit To Post
Is Maradona 'Messing' with Messi?
Good Witch / Bad Witch
High School Reunion: The Potentially Non-Musical
Perusals & Peregrinations
Words is Me

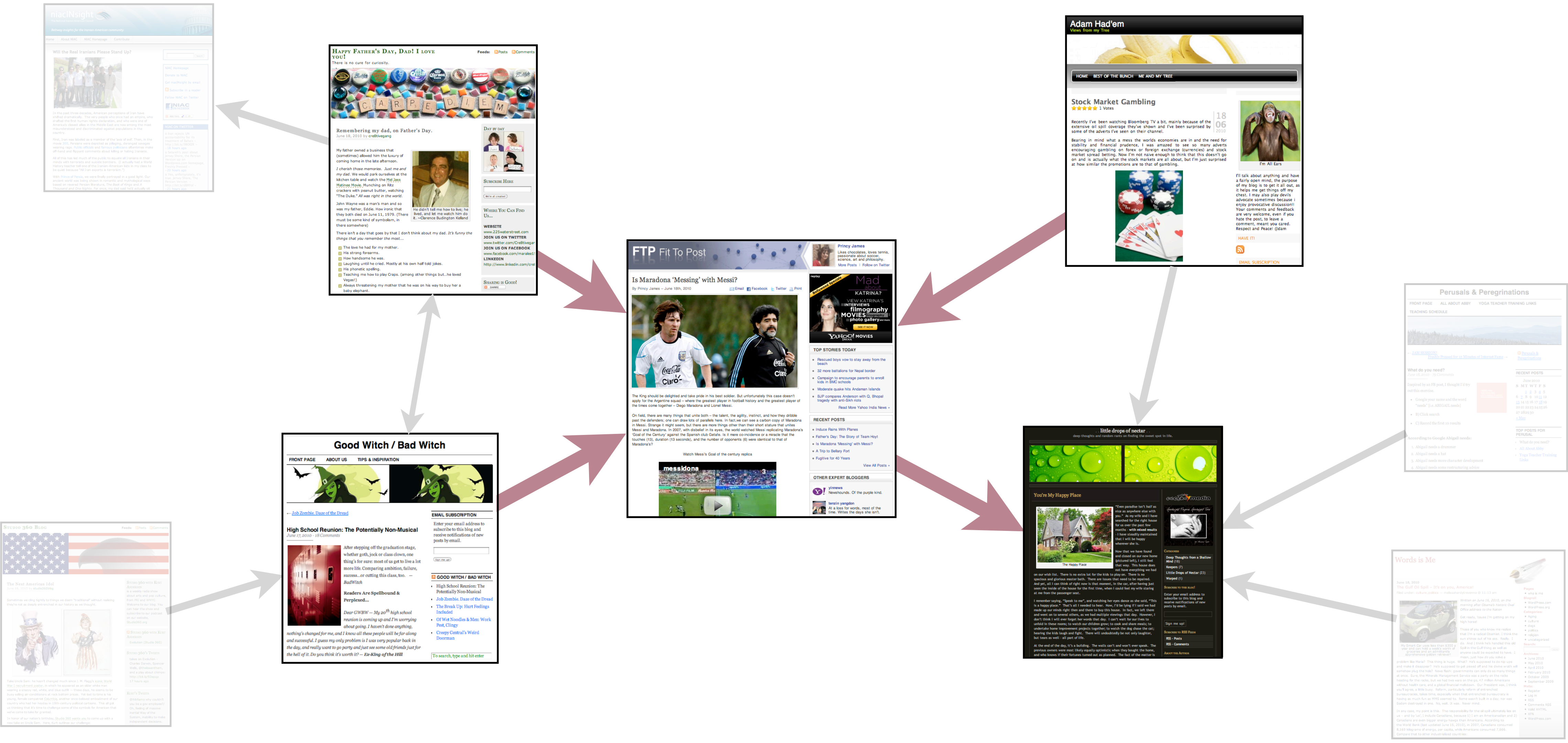
In-degree Centrality = 3

Network Analysis


Happy Father's Day, Dad! I love you!
Adam Had'em
Stock Market Gambling
FTP Fit To Post
Is Maradona 'Messing' with Messi?
Perusals & Peregrinations
Good Witch / Bad Witch
High School Reunion: The Potentially Non-Musical
little drops of nectar
You're My Happy Place
Words is Me

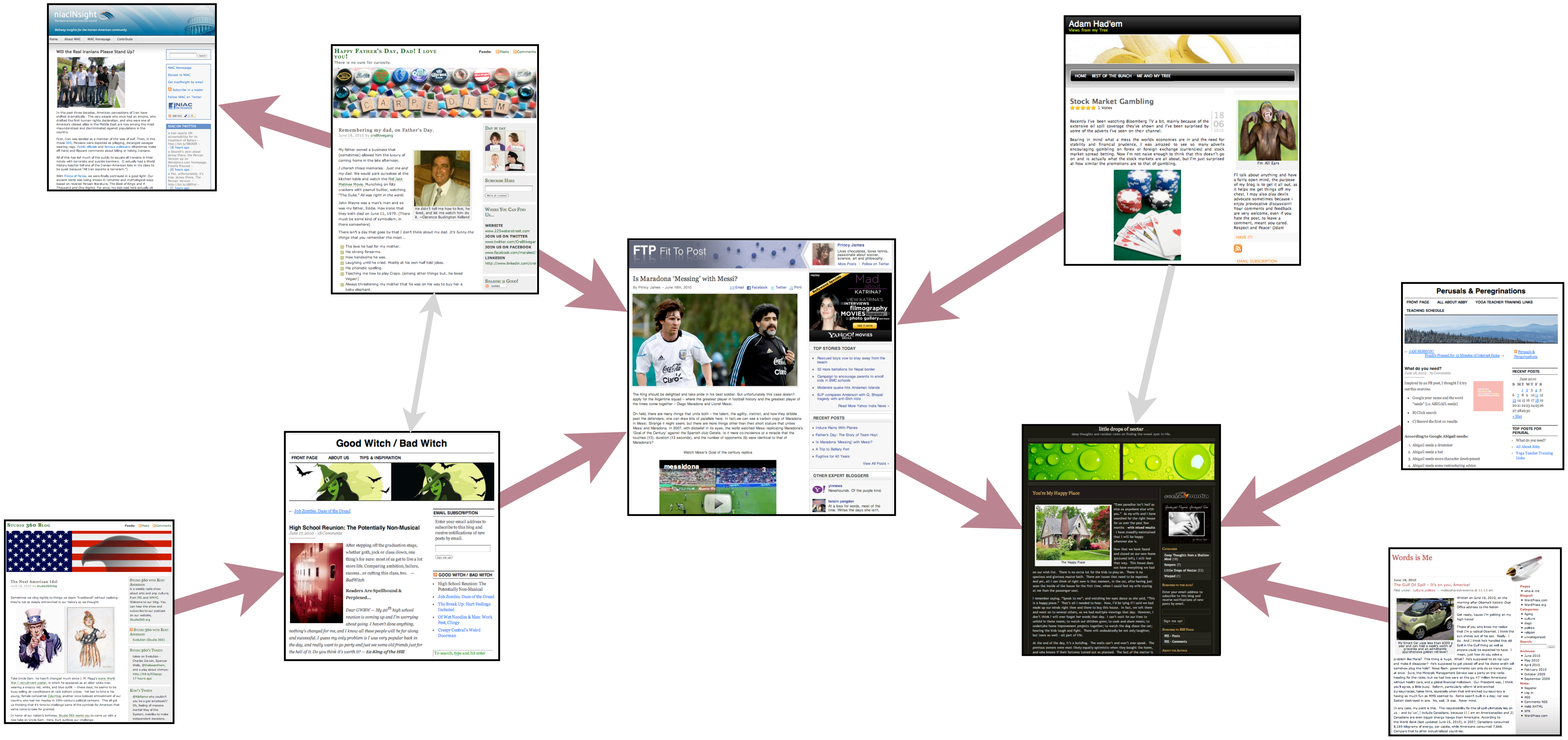
Out-degree Centrality = 1

Network Analysis


Happy Father's Day, Dad! I love you!
FTP Fit To Post
Is Maradona 'Messing' with Messi?
Adam Had'em
Stock Market Gambling
Good Witch / Bad Witch
High School Reunion: The Potentially Non-Musical
little drops of nectar
You're My Happy Place
Perusals & Peregrinations
Words is Me

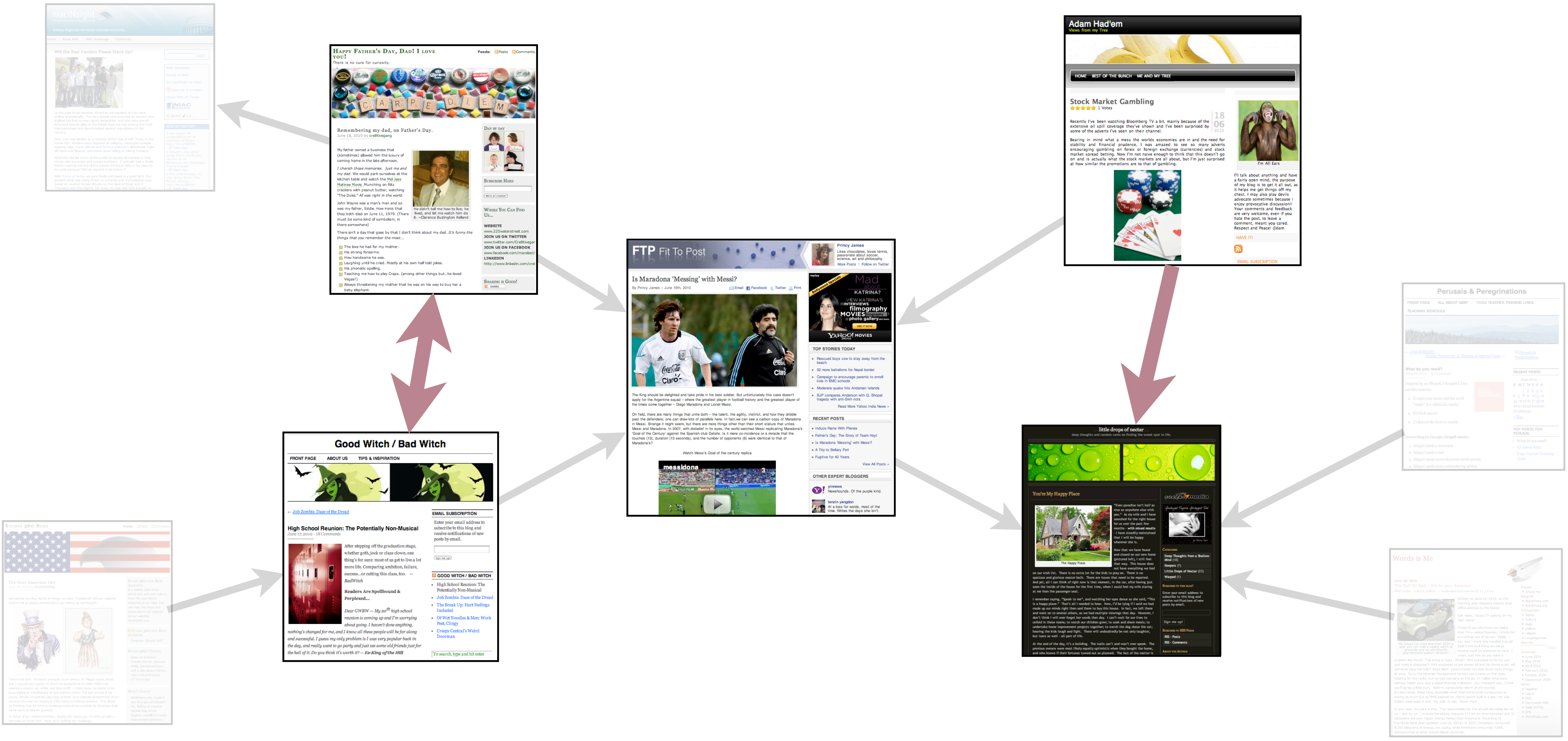
Total-degree Centrality = 4

Network Analysis



Betweenness Centrality = 16

Network Analysis


Happy Father's Day, Dad! I love you!
Adam Had'em
FTP Fit To Post
Is Maradona 'Messing' with Messi?
Good Witch / Bad Witch
High School Reunion: The Potentially Non-Musical
little drops of nectar
Stock Market Gambling
Perusals & Peregrinations
Words is Me

Clustering Coefficient = 1/3

“Influence Size”

Takes into account

- how many readers write posts in response
- how similar their topics are

$$S_i = \{B_j : \ B_j \text{ can reach } B_i \text{ by following links}\}$$

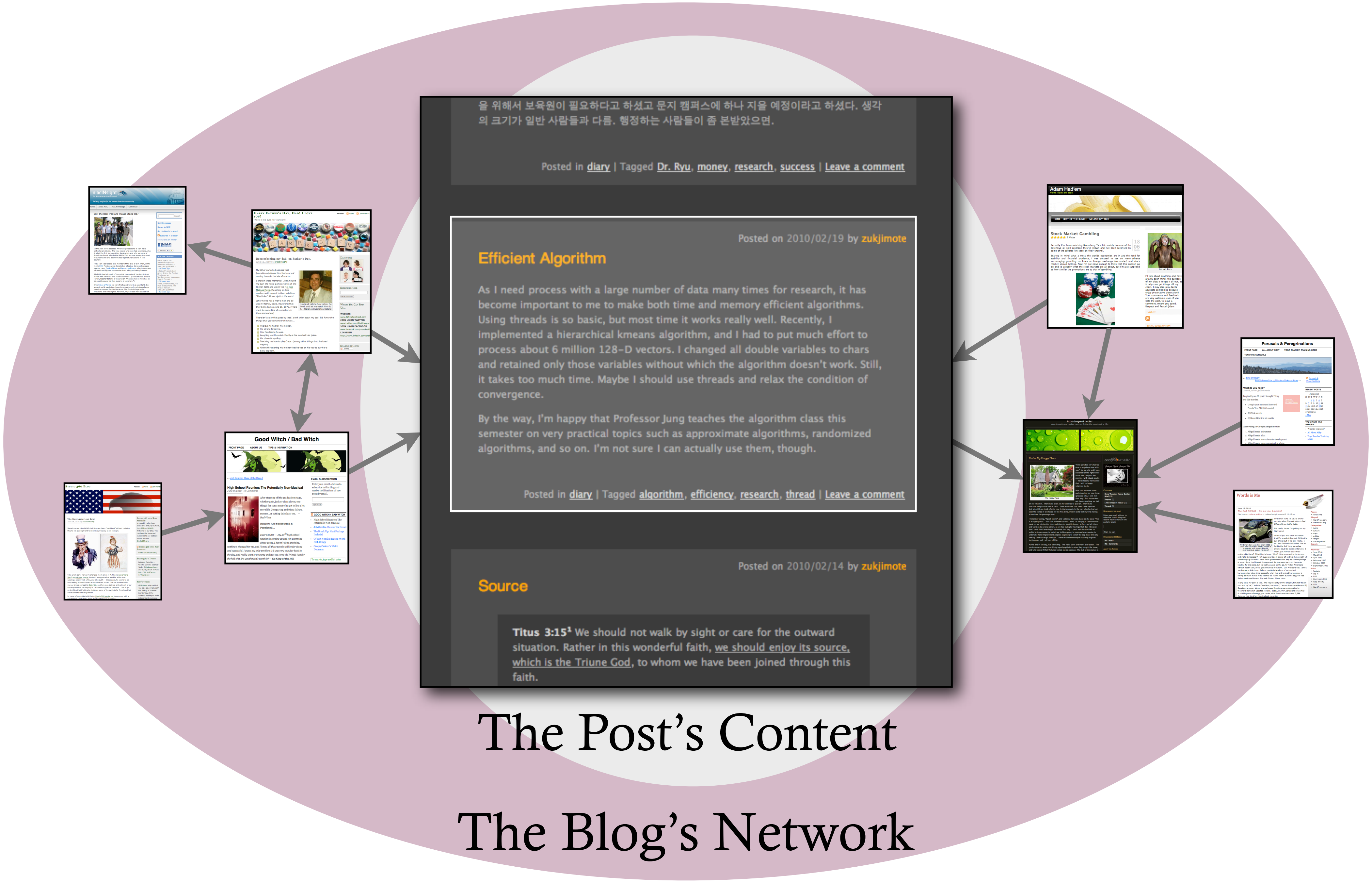
$$\begin{aligned}\text{InfluenceSize}(B_i) &= \sum_{B_j \in S_i} \text{TopicSimilarity}(B_i, B_j) \\ &= \sum_{B_j \in S_i} \frac{\mathbf{t}_i \cdot \mathbf{t}_j}{\|\mathbf{t}_i\| \|\mathbf{t}_j\|}\end{aligned}$$

Existing Measures of Blog Influence

- Number of Comments
  - Quantitative
- Digg Score (Digg.com)
  - Quantitative, partly qualitative

> “How can we predict the potential influence of a new blog post?”

Prediction of Blog Influence



을 위해서 보육원이 필요하다고 하셨고 문지 캠퍼스에 하나 지을 예정이라고 하셨다. 생각의 크기가 일반 사람들과 다름. 행정하는 사람들이 좀 본받았으면.

Posted in diary | Tagged Dr. Ryu, money, research, success | Leave a comment

Posted on 2010/02/19 by zukjimote

Efficient Algorithm

As I need process with a large number of data many times for research, it has become important to me to make both time and memory efficient algorithms. Using threads is so basic, but most time it works really well. Recently, I implemented a hierarchical kmeans algorithm, and I had to put much effort to process about 6 million 128-D vectors. I changed all double variables to chars and retained only those variables without which the algorithm doesn't work. Still, it takes too much time. Maybe I should use threads and relax the condition of convergence.

By the way, I'm happy that Professor Jung teaches the algorithm class this semester on very practical topics such as approximate algorithms, randomized algorithms, and so on. I'm not sure I can actually use them, though.

Posted in diary | Tagged algorithm, efficiency, research, thread | Leave a comment

Posted on 2010/02/14 by zukjimote

Source

Titus 3:15[1] We should not walk by sight or care for the outward situation. Rather in this wonderful faith, we should enjoy its source, which is the Triune God, to whom we have been joined through this faith.

The Post's Content

The Blog's Network

Experiments

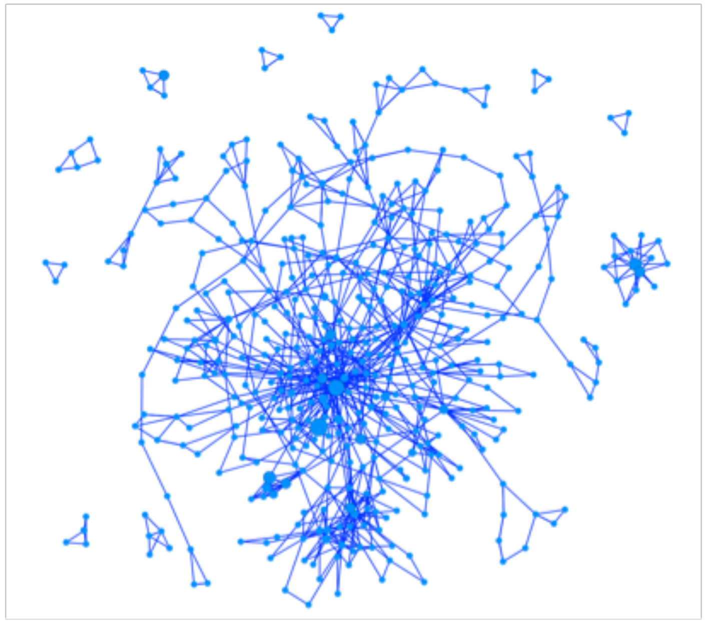
1. Content information and network information capture different aspects of blog influence
2. It is important to use both content information and network information for finding influential blogs

Dataset

- Selected from TREC Blog08
  http://ir.dcs.gla.ac.uk/test_collections/blogs08info.html

# of Blogs	# of Posts	# of Unique Words	Average # of Words/Post
4,165	72,143	53,257	225.24

- ✓ Blogspot, LiveJournal
- ✓ Contains >50 words
- ✓ Has at least one link to another post
- ✓ Written in 2008
- ✓ Written in English

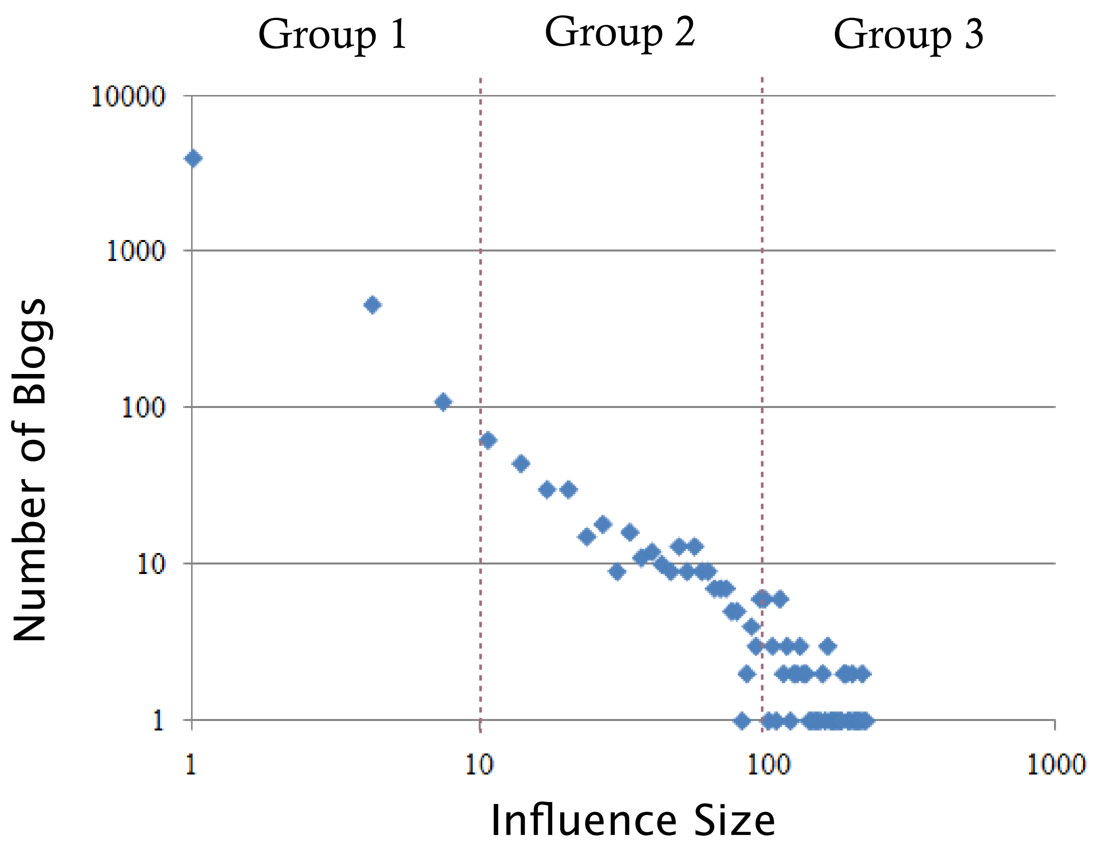


Experiment I

“Do content information and network information play different roles in predicting blog influence?”

Design

- **Task**
  Classify the blogs into three groups
  (non-influential, influential, and very-influential)
- **Ground Truth**
  Sort the blogs in the order of *influence size* and group them into three



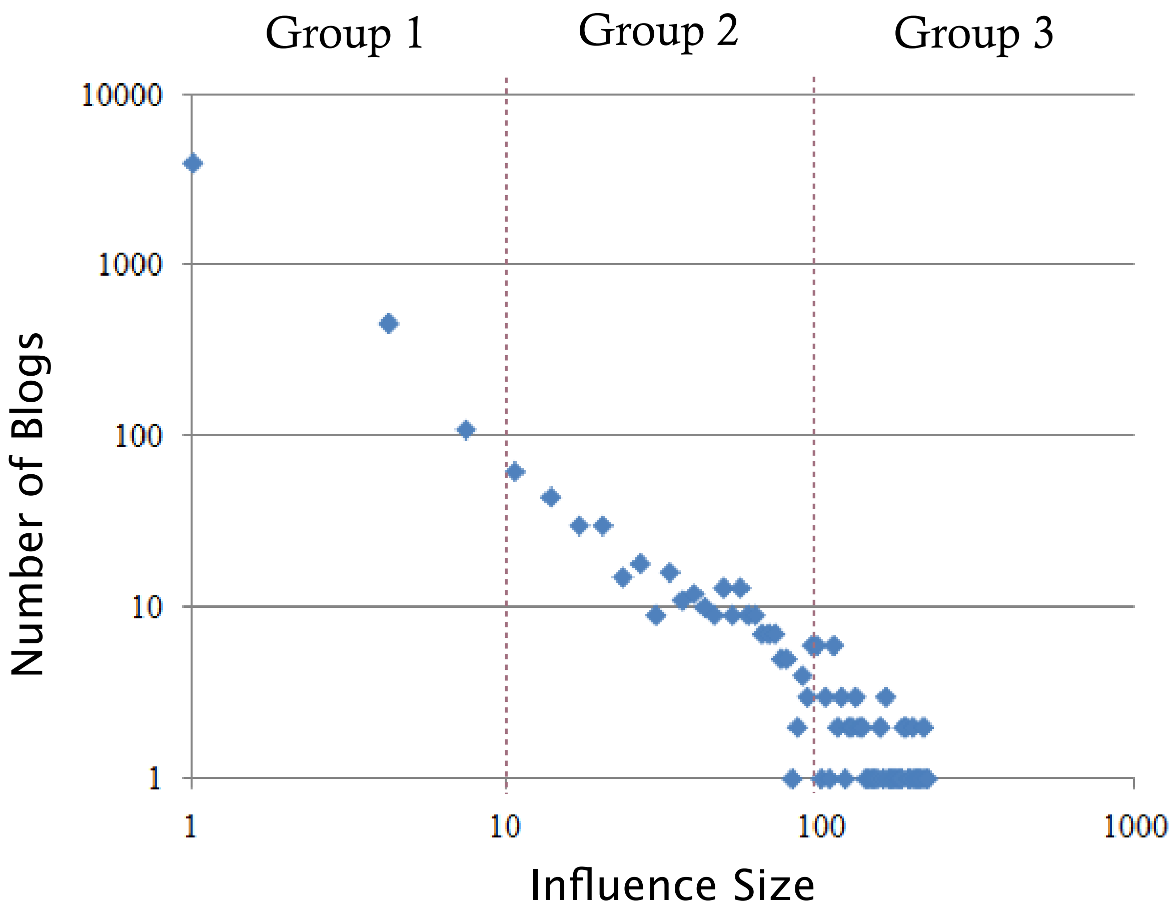
Classification

- **Linear SVM**
- **Features**
  - Content: Topic proportions (50)
  - Network: In-degree, out-degree, total-degree, betweenness, clustering coefficient (5)
- **Training**

  Train a classifier using 30% of the data blogs
- **Testing**

  Test with the rest of the data

Feature Importance



Group 1	Group 2	Group 3
OutDegree	OutDegree	Topic 47
TotalDegree	TotalDegree	Topic 14
Betweenness	Topic 9	Topic 15
Clustering Coefficient	Topic 35	Topic 43
Topic 8	Topic 13	Topic 41
Topic 28	Topic 0	Topic 38
Topic 19	Topic 24	Topic 42
InDegree	Topic 27	Topic 24
Topic 43	Topic 16	Topic 12
Topic 23	Topic 19	Topic 22

Most Important Features

Experiment II

> “If we use both content information and network information, can we find potentially influential blogs better than when using only one of them?”

Design

- **Task**

  Classify each blog whether it belongs to the top 10% influential blogs or not

- **Ground Truth**

  Top 10% influential blogs are obtained with regard to each influence measure (influence size, number of comments, and Digg score)

Classification

- **Linear SVM**
- **Features**
  - Content: Topic proportions (50)
  - Network: In-degree, out-degree, total-degree, betweenness, clustering coefficient (5)
  - Content-Network: All (55)
- **Training**
  Train a classifier using 30% of the data blogs
- **Testing**
  Test with the rest of the data

Prediction Result

Features	Influence Size	Number of Comments	Digg Score
Content	0.360	0.275	0.295
Network	0.726	0.308	0.239
Content-Network	**0.727**	**0.322**	**0.308**

F1-Measure

(Harmonic mean of precision and recall)

Prediction Result

Features	Influence Size				Number of Comments				Digg Score			
	Accuracy	Precision	Recall	F1-Measure	Accuracy	Precision	Recall	F1-Measure	Accuracy	Precision	Recall	F1-Measure
Content	0.841	0.303	0.445	0.360	0.631	0.172	0.691	0.275	0.688	0.197	0.583	0.295
Network	0.934	0.683	0.775	0.726	0.857	0.307	0.309	0.308	0.851	0.283	0.207	0.239
Content-Network	0.943	0.699	0.759	0.727	0.716	0.213	0.664	0.322	0.714	0.211	0.571	0.308

Contributions

- We proposed a new measure of blog influence, *Influence Size*, which considers both content and network
- We showed that content information and network information play different roles in predicting blog influence
- We showed that it is important to use both content information and network information for finding influential blogs

Future Work

- Evaluation of the influence measures
- Exploration of various methods for content analysis

Thank You

Presenter

Yohan Jo
yohan.jo@kaist.ac.kr
Users & Information Lab
Computer Science Dept.
Korea Advanced Institute of Science and Technology