Identifying a Topic Lifecycle

- Research Objectives
- Research Plan I (Kick-off meeting in Warsaw)
- System Implementation (Part I)
- Experimental Procedure
- The Blog Data Used
- Experimental Result

- Research Plan II (Project meeting in Karlsruhe)
- Identify a topic lifecycle
- Extension to the Existing System Implementation
- Summary

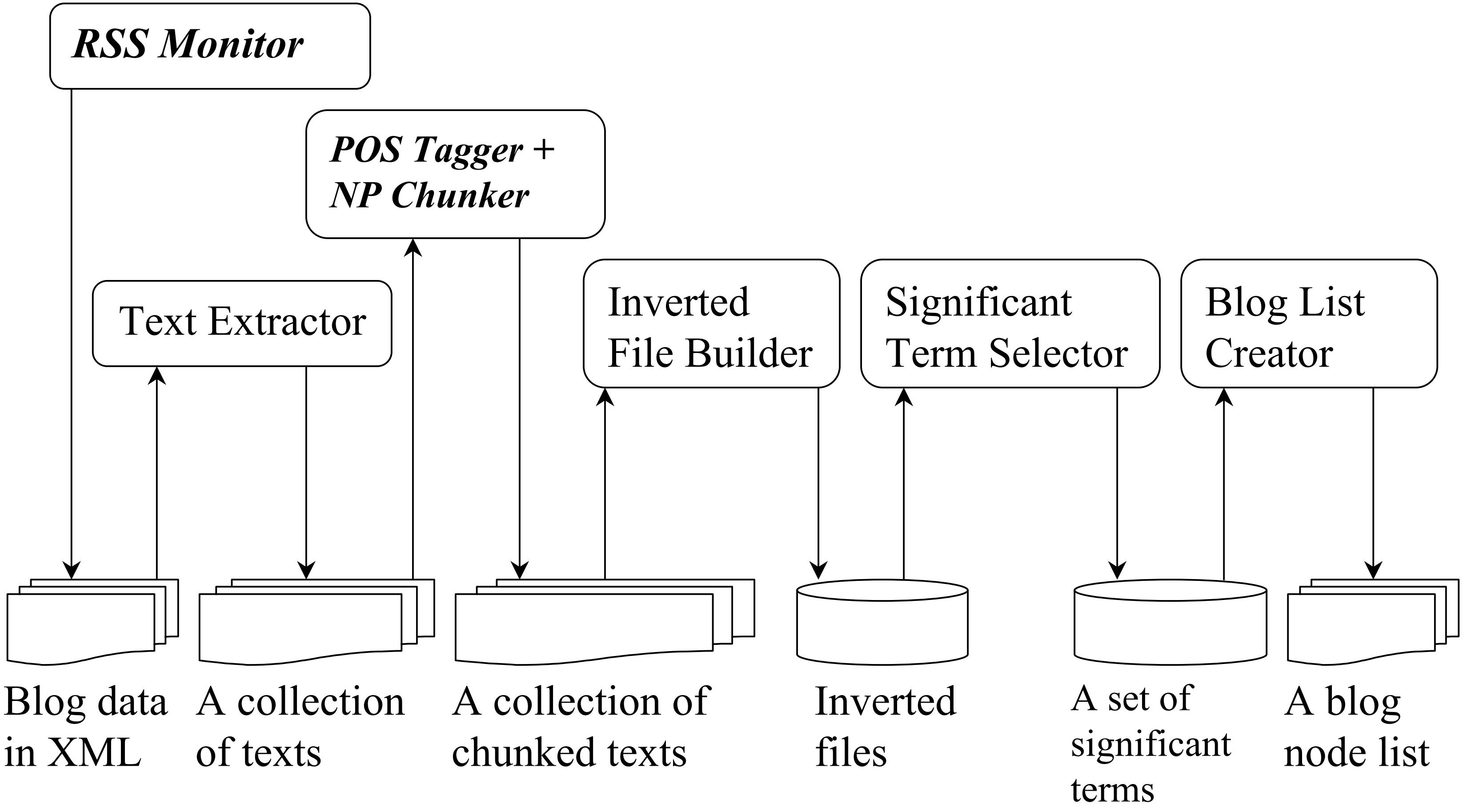
Research Objectives

- Identify emerging topics about public science debates (*topic identification*), such as GM foods and foot-and-mouth disease;
- Keep track of the flow/spreading of the emerging topics (*topic tracking*).

Research Plan I

- Extract a set of terms from a collection of texts.
- Apply a technique to select a set of significant terms which may represent emerging topics about science debates.
- Group the significant terms into a number of blogs.

System Implementation (Part I)



Experimental Procedure

- Apply three term selection methods independently, in order to select a set of significant terms on a certain date. These are chi-square statistics, mutual information and information gain.
  - Build 2x2 contingency table, so that chi-square and mutual information values can be computed.
  - Build a table which stores the entropy value of each date, so that information gain can be computed
- Select a number of topics manually as a starting point.
- Select a number of significant terms with respect to each topic automatically.
- Group the significant terms into a number of blogs.

The Blog Data Used

#Blog	#Raw data	#Item	#Term	#Inverted file	#term /item
3,702	50 MB	117,652	264,994	1,076,152	4.06
19,587	350 MB	880,536	1,736,715	8,436,624	4.86
19,587	1.34 GB	-	-	-	-

Experimental Results

#Item	#Term	#Topic	#Blog node
117,652	264,994	36	127
880,536	1,736, 715	44	2157

Top 10 Topics	#Tuple
tsunami	2037
aids + hiv	799
pollution	755
agriculture	568
bird-flu	494
breast-cancer	433
climate-chage	326
terrorism + bioterrorism	282
global-warming	257
diabetes	255

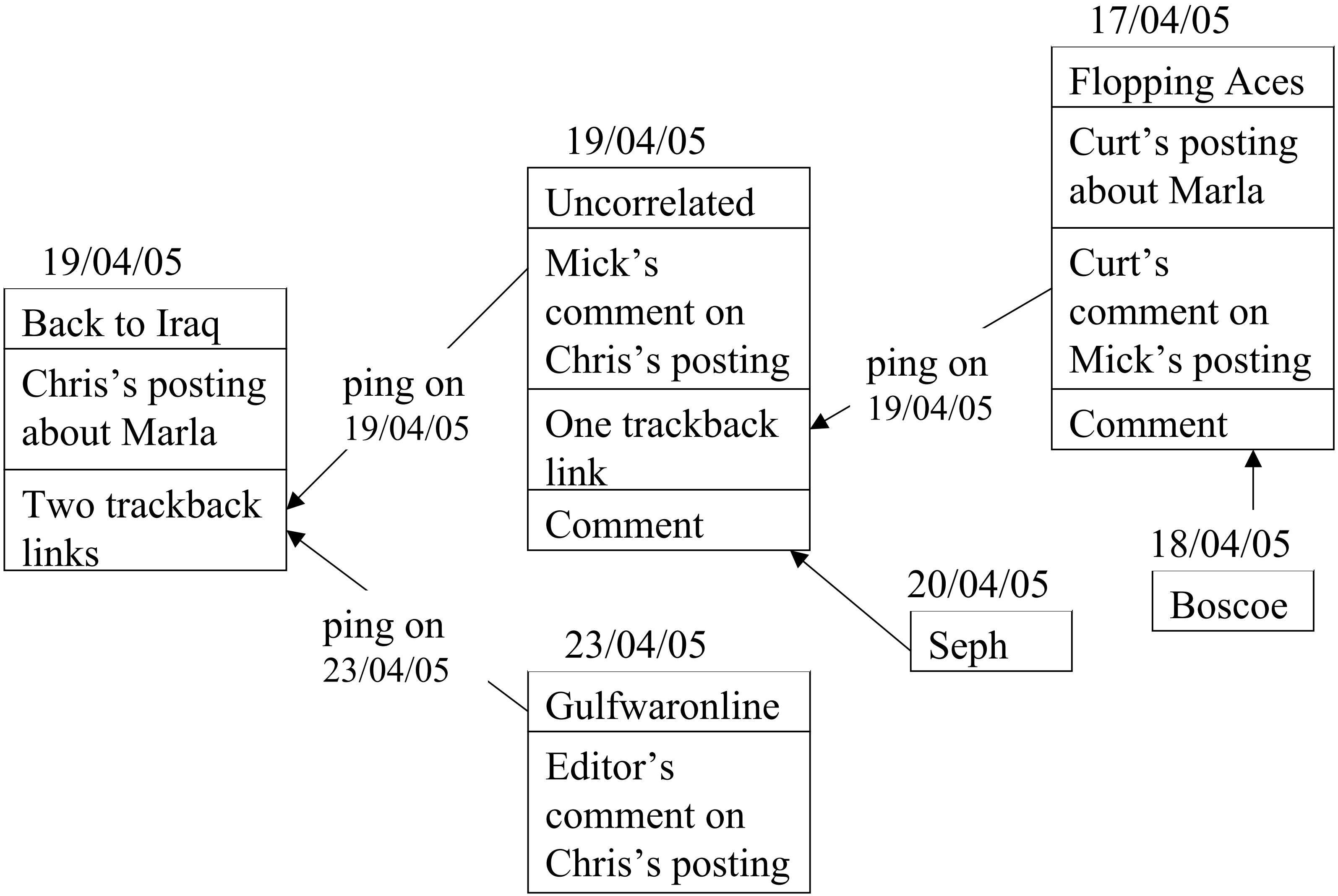
Research Plan II

- Identify a set of emerging topics
- Track the lifecycle of the emerging topics, from the time they were born, until the time they are fading away.
- Two approaches have been considered.
  - Link-based approach
  - Term-based approach

Identifying a Topic Lifecycle: a Link-Based Approach

- Existing type of links
  - Permalink (= Permanent link).
  - Trackback link (= Tracking the person who makes a comment on a particular posting (or item)).
  - Blogrolls (= a collection of blog URLs within a blog).
- Trackback links are suitable for link analysis, but there are not too many trackback links on the Blog sites available.
- Blogrolls should not be used for tracking a certain topic without the use of trackback links.

A Link-Based Scenario about Marla Ruzicka



19/04/05

Back to Iraq

Chris's posting about Marla

Two trackback links

ping on 19/04/05

ping on 23/04/05

19/04/05

Uncorrelated

Mick's comment on Chris's posting

One trackback link

Comment

23/04/05

Gulfwaronline

Editor's comment on Chris's posting

ping on 19/04/05

20/04/05

Seph

17/04/05

Flopping Aces

Curt's posting about Marla

Curt's comment on Mick's posting

Comment

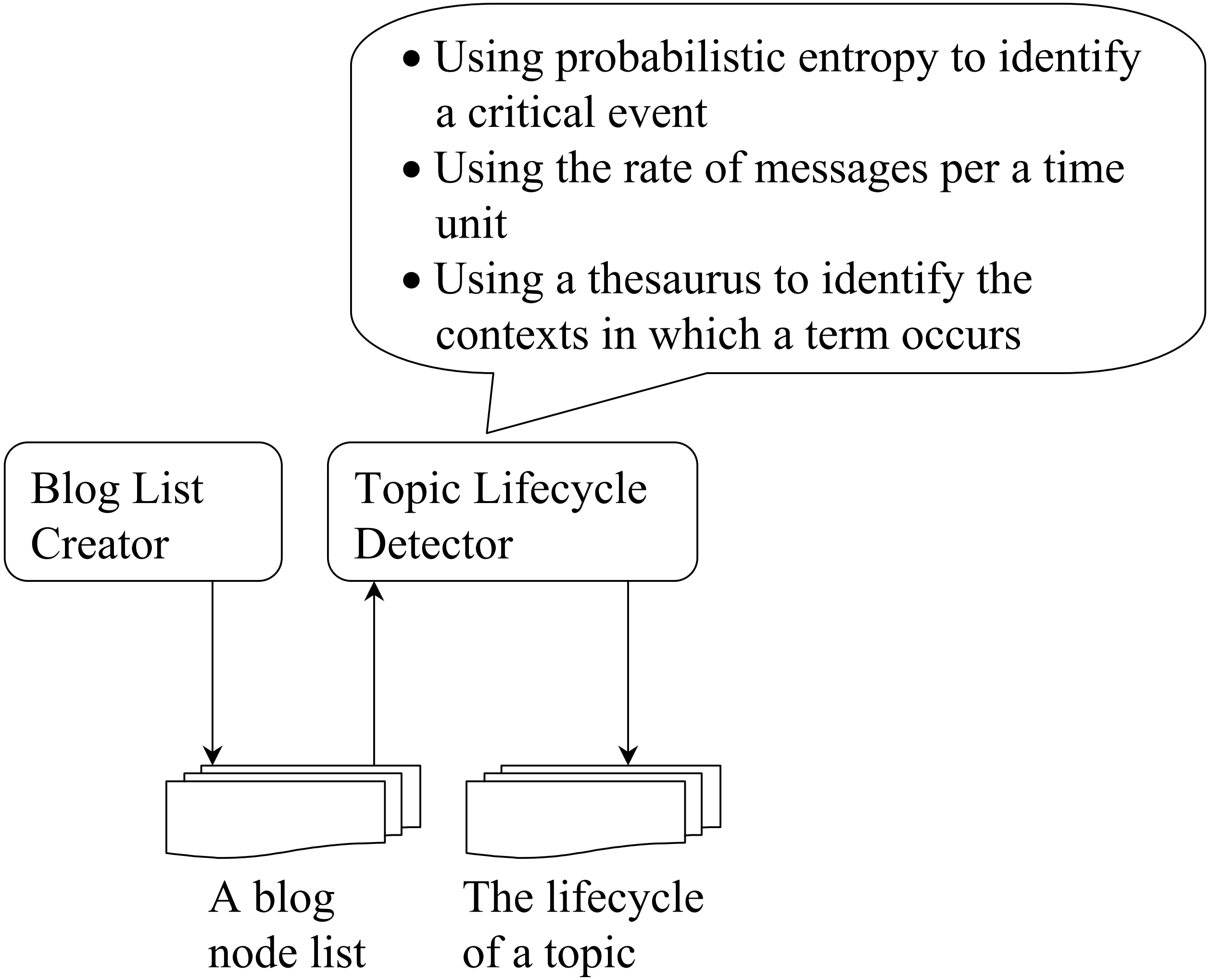
18/04/05

Boscoe

Identifying a Topic Lifecycle: a Term-Based Approach

- Using probabilistic entropy, the one Loet describes in his email, to identify a critical event.
- Using Kleinberg’s topic burst detection algorithm, which is based on the rate of messages per time unit (e.g. per hour/per day).
- Using a thesaurus to identify the relationship between terms, such as WordNet and Wikipedia to analyse the contexts in which a term occurs, e.g. the nouns / modifiers which co-occur with the term.

Extension to the Existing System Implementation



- Using probabilistic entropy to identify a critical event
- Using the rate of messages per a time unit
- Using a thesaurus to identify the contexts in which a term occurs

The Construction of a Network Structure of Topic Spreading between Bloggers: the Assumptions

- The occurrence of a phrase within two blog nodes, i.e. a source node and a target node, on a consecutive day may indicate that there is a link between the two blog nodes. It is especially significant if the number of edges between the two blog nodes are greater than a threshold value.
- The source and target nodes are determined based on the phrase publication date, where

source_node.pub_date(phrase) < target_node.pub_date(phrase)

- All self-links are discarded, since they do not represent the spreading of a topic.
- Then, an undirected graph is constructed.

The Construction of a Network Structure of Topic Spreading between Bloggers: an Illustration (#node = 2,157 & #edge = 12,069)

Input :

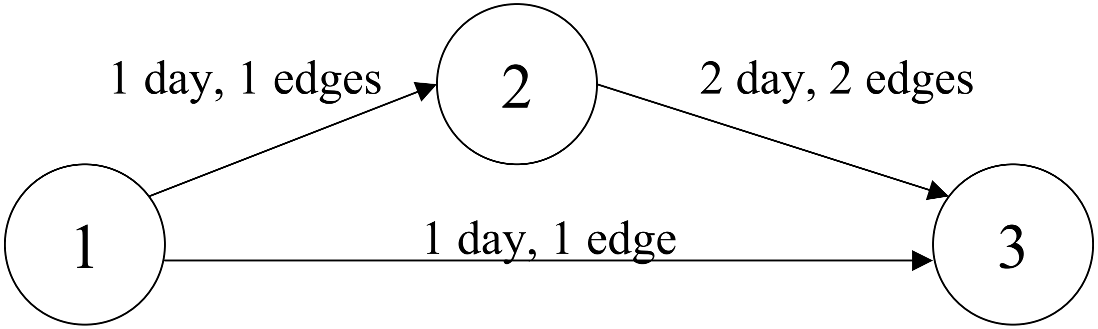
CREEN meeting

Date	Blog
22/12/2004	1
23/12/2004	1 , 2
24/12/2004	1 , 2
25/12/2004	3

Knowledge Domain Visualisation

Date	Blog
23/12/2004	2
25/12/2004	3

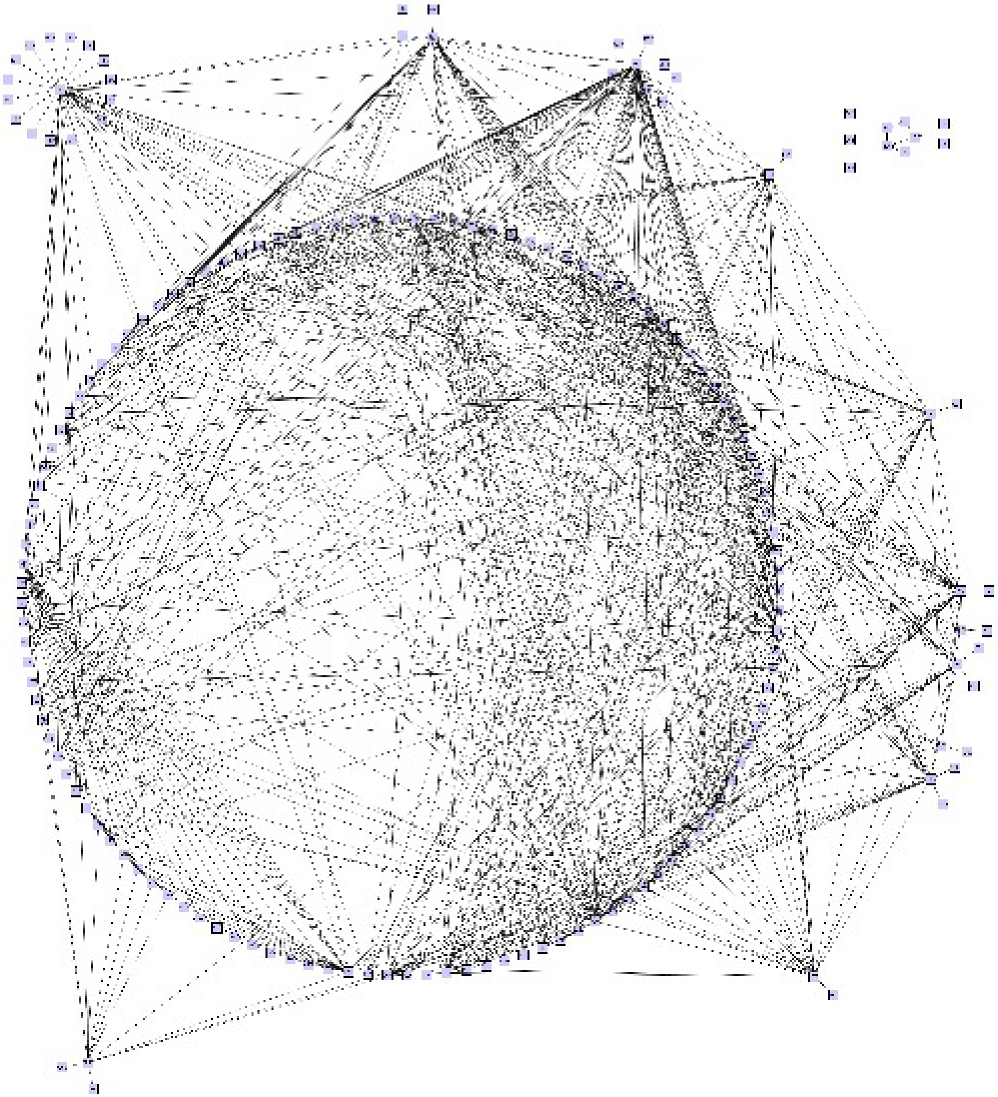
Output :



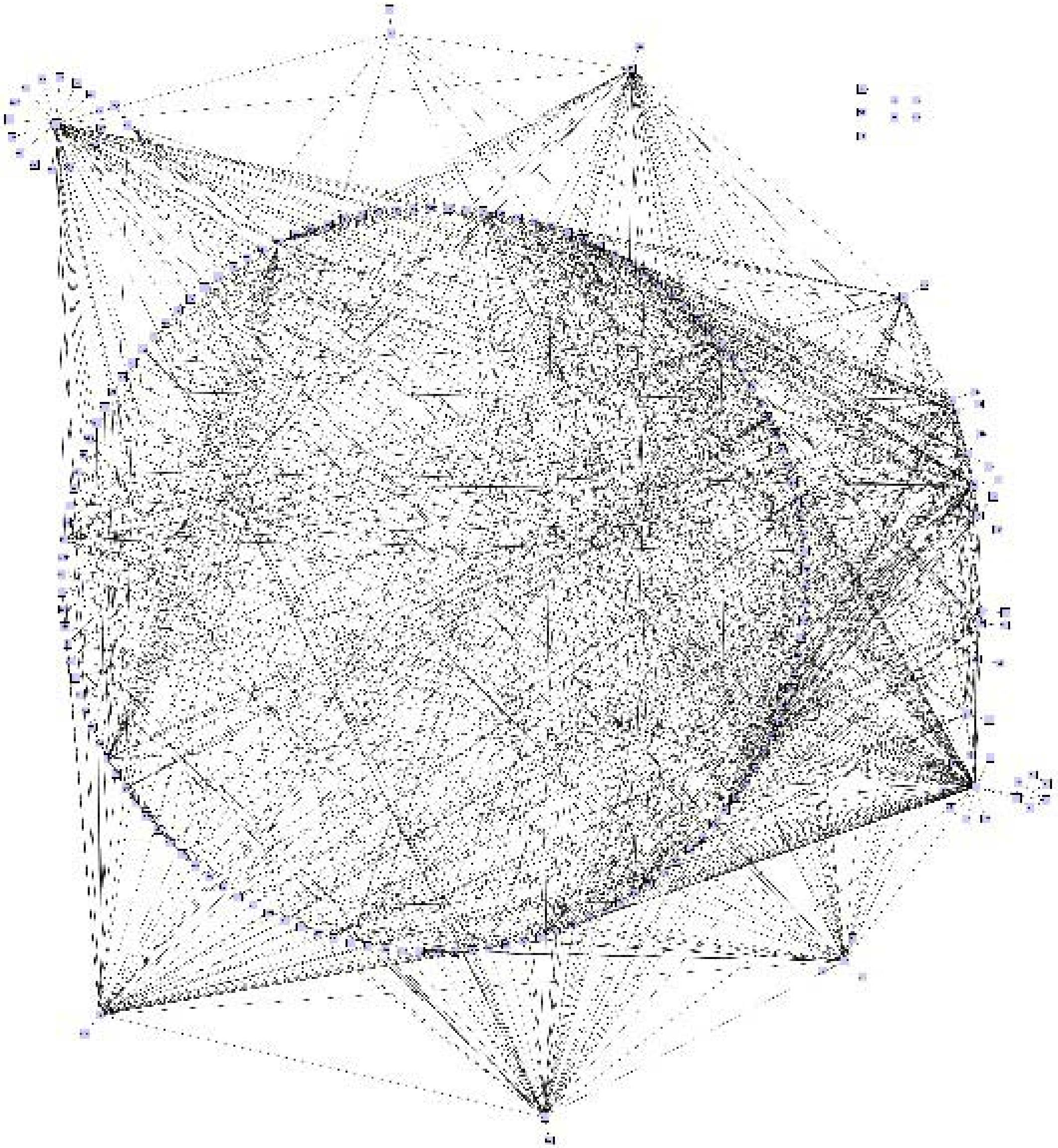
The Construction of a Network Structure of Topic Spreading between Bloggers: the Issues

- Either blog rolls or the occurrence of a phrase within two blog nodes cannot be used to draw a conclusion, but to indicate that there is a connection between the two nodes.
- Trackbacks offer a better mechanism to show the connection between two nodes, but they are quite sparse. Thus, the coverage level of a network structure can become an issue. Given 100 blogs, 11 blogs contain trackback labels; 5 of which contain trackback links.
- Combined approach: Blog rolls + the occurrence of a phrase + trackbacks ?

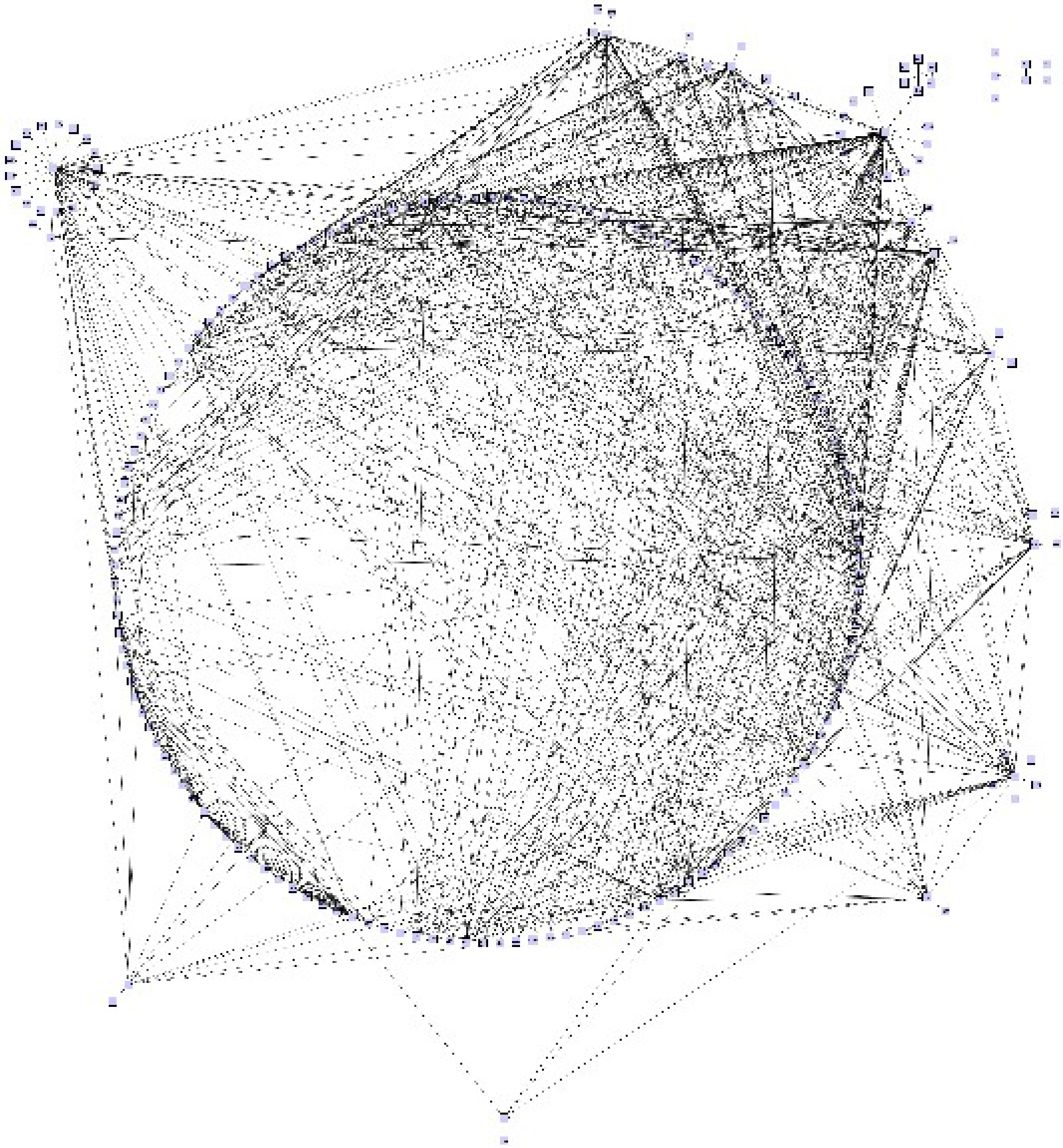
The Visualisation of Topic Spreading between Bloggers (time_period <= 1 day & min_num_of_edges = 3)



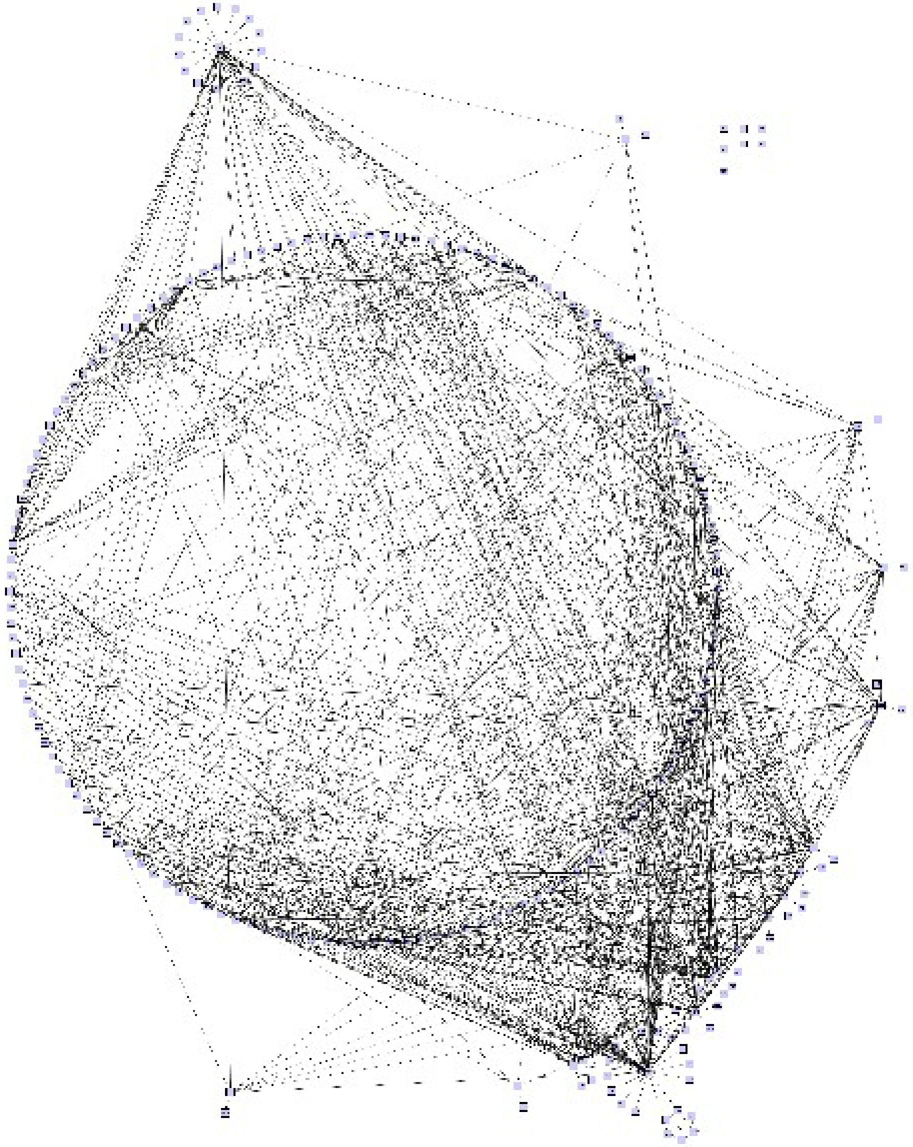
The Visualisation of Topic Spreading between Bloggers (time_period <= 7 days & min_num_of_edges = 3)



The Visualisation of Topic Spreading between Bloggers (time_period <= 14 days & min_num_of_edges = 3)

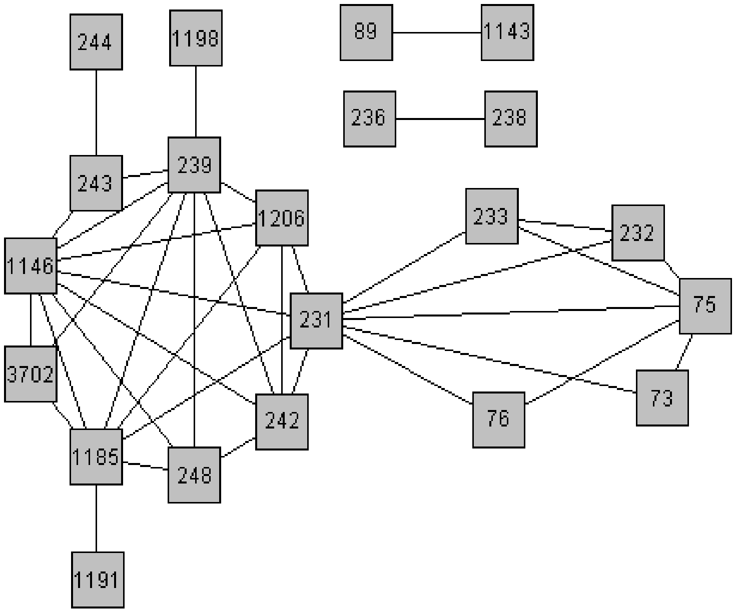


The Visualisation of Topic Spreading between Bloggers (time_period <= 1 year & min_num_of_edges = 3)

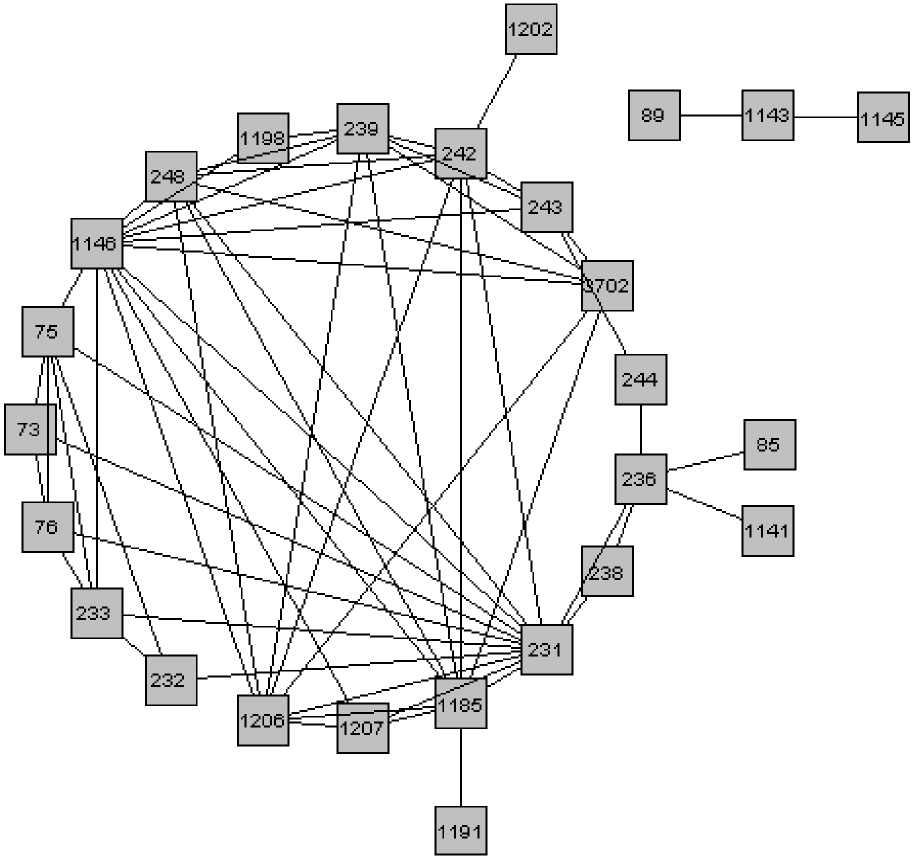


The Visualisation of Topic Spreading between Bloggers (min_num_of_edges = 25)

time_period >= 1 day

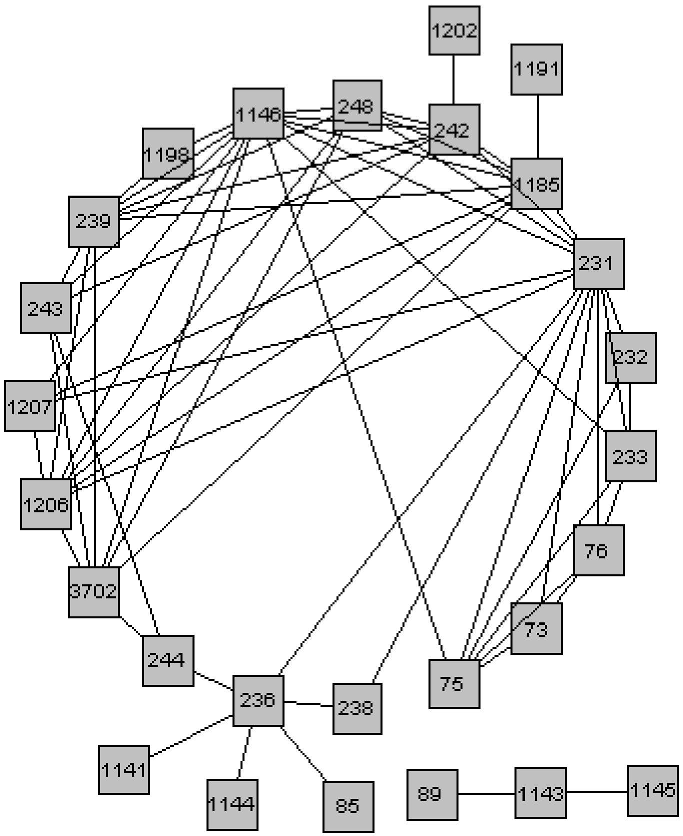


time_period >= 7 days

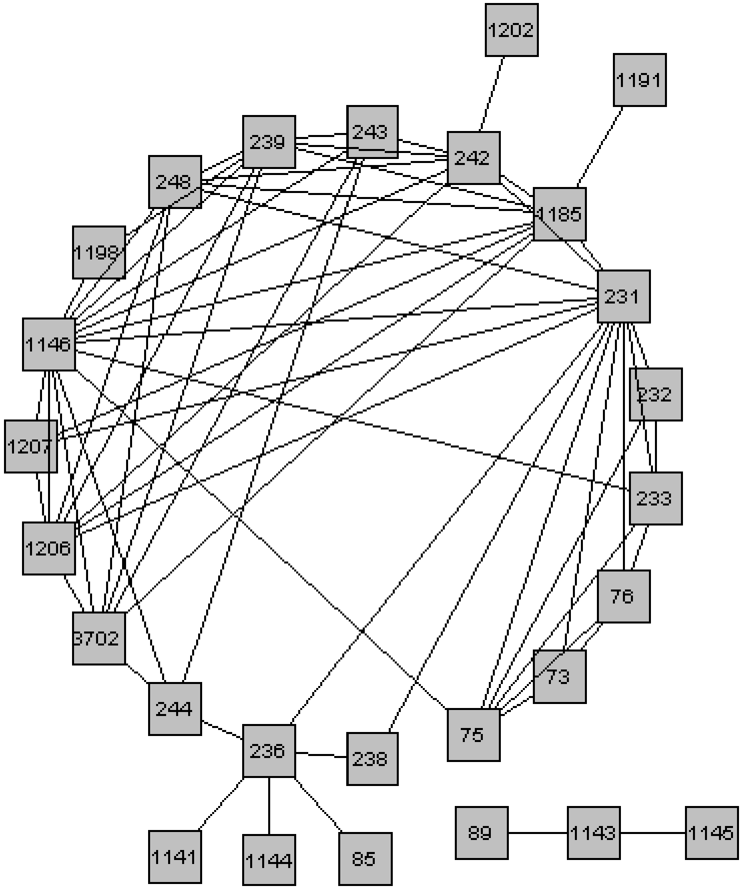


The Visualisation of Topic Spreading between Bloggers (min_num_of_edges = 25)

time_period >= 14 days



time_period >= 1 year



#Node & #Edge

min_num_of_edges = 3

Time period (in days)	#Node	#Edge
1	**179**	**717**
7	**207**	**855**
14	**220**	**893**
30	**224**	**905**
60	**224**	**913**
90	224	914
120	224	914
150	225	915
180	225	915
210	225	915
240	225	915
270	225	915
300	225	915
330	225	915
365	225	916

min_num_of_edges = 25

Time period (in days)	#Node	#Edge
1	**21**	**37**
7	**26**	**60**
14	27	62
30	27	63
60	27	63
90	27	63
120	27	63
150	27	63
180	27	63
210	27	63
240	27	63
270	27	63
300	27	63
330	27	63
365	27	63

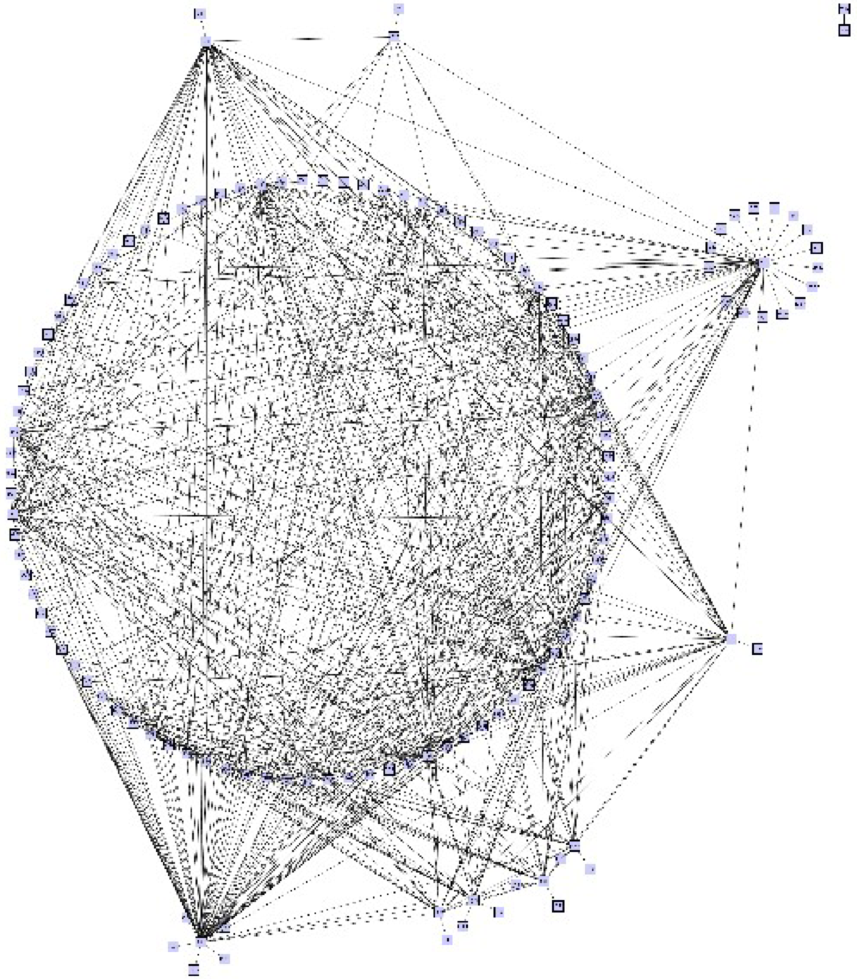
Edge Frequency

Edge-Freq	Count
1	9601
2	1554
3	336
4	164
5	98
6	51
7	36
8	27
9	22
10	8
11	16
12	22
13	11
14	9
15	6
16	4
17	11
18	7
19	2
20	3
21	6
22	4
23	4
24	4
25	2
26	2
27	2
29	1
30	5
31	5
32	2
34	2
35	2
36	2
37	2
38	1
39	3
41	2
42	1
43	1
44	4
45	2
46	2
49	2
50	1
51	1
52	1
57	1
61	1
69	1
71	2
76	1
77	1
79	1
80	1
81	1
90	1
95	2
98	1
104	1

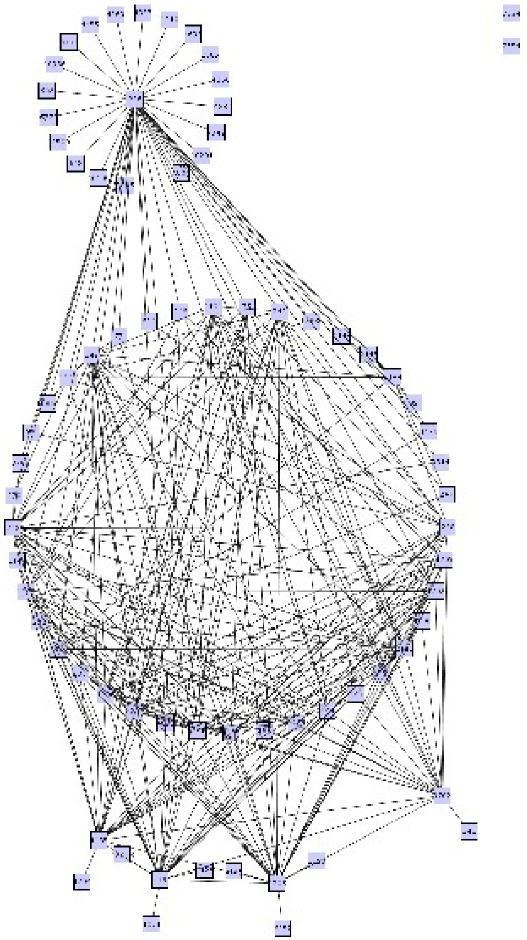
Indications I

- Bloggers are interested in writing something, or commenting on other blogger postings over a period of time, e.g. within 7 days until 3 months.
- They compose a cluster which may represent their common interest in a certain topic.
- The smaller the edge frequency, the larger the number of the edges w.r.t the edge frequency.
- The number of edges which have the largest edge frequency are the smallest. This indicates that only few bloggers have a strong communication between them.

The Visualisation of Topic Spreading between Bloggers (01/12/2004 – 31/01/2005)

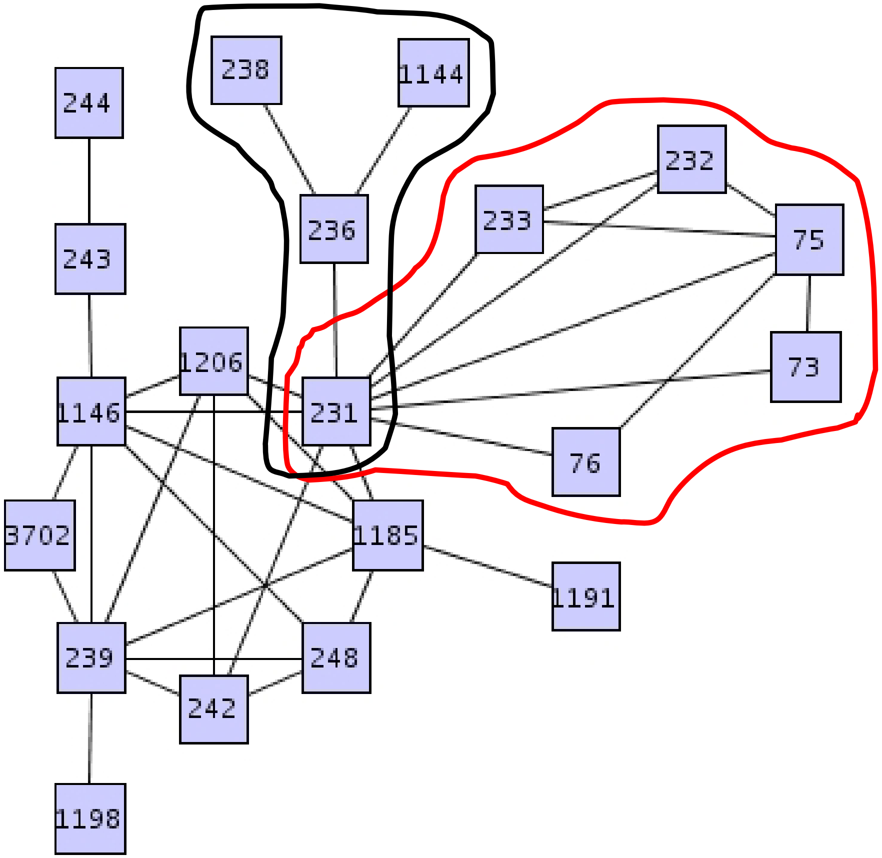


<= 1 day; 3 edges; with tsunami
#node = 133
#edge = 598

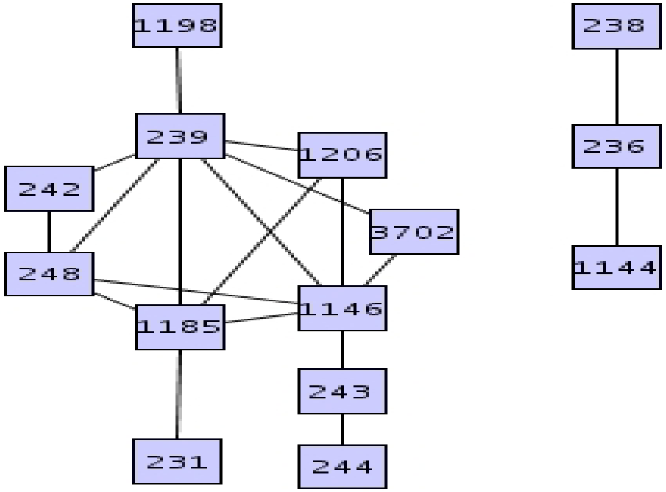


<=1 day; 3 edges; without tsunami
#node = 75 (43.6% reduction)
#edge = 294 (50.84% reduction)

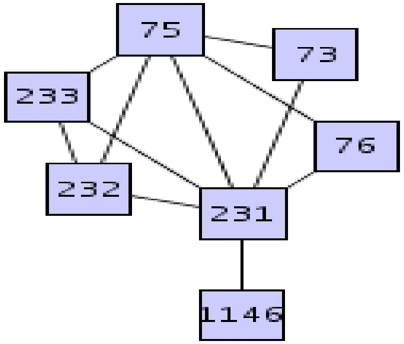
The Visualisation of Topic Spreading between Bloggers
(01/12/2004 – 31/01/2005), <=1 day, min_num_of_edges = 25 edges



with tsunami: #node = 20, #edge = 35

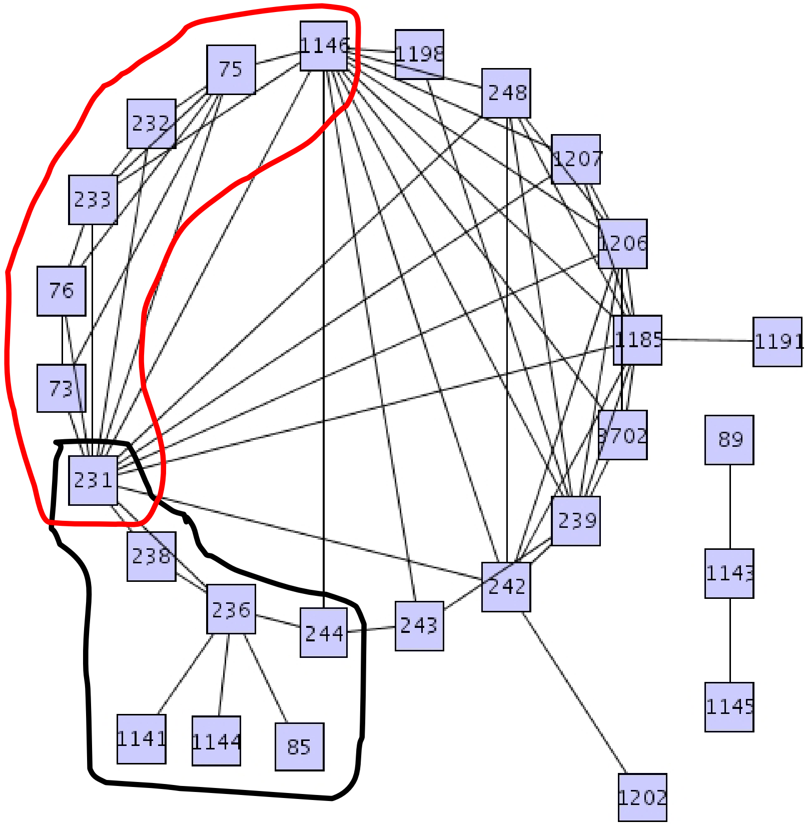


without tsunami:
#node = 14, #edge = 19

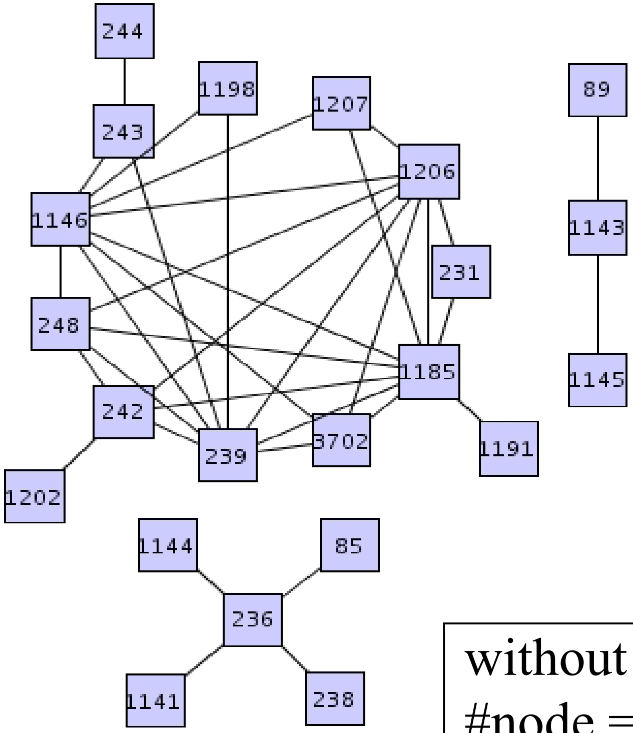


only tsunami:
#node = 7, #edge = 11

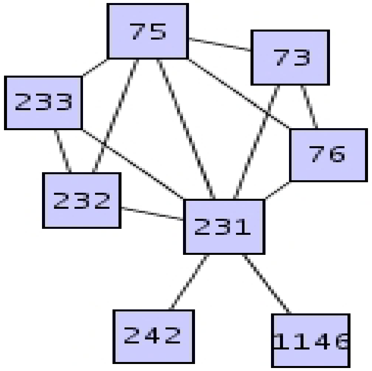
The Visualisation of Topic Spreading between Bloggers (01/12/2004 – 31/01/2005), <= 14 days, min_num_of_edges = 25 edges



with tsunami:
#node = 27 , #edge = 59



without tsunami:
#node = 22, #edge = 36



only tsunami:
#node = 8, #edge = 13

Text Document

Indications II

- The occurrence of a significant event, such as Tsunami, changes the structure of a blogosphere.
- An existing node can enter the main circle.
- An existing cluster may connect to the main circle.
- Some nodes can increase their edge frequency significantly, create a new cluster, and connect to the main circle

Summary

- A collection of blog data were processed to extract a number of significant terms on a certain date.
- The output is a blog node list which contains a number of significant terms - for each blog - on a certain date.
- Based on the blog node list, the lifecycle of a topic will be identified.
- Two approaches have been considered: link-based and term-based approaches.
- Since link-data is sparse, it is likely that the term-based approach will be used to identify the lifecycle of a topic, or the combination of both approaches.
- The occurrence of a significant event changes the structure of a blogosphere.