CS 5604 - Final Presentation

CINETGraphCrawl

Rushi Kaw
Hemanth Makkapati
Rajesh Subbiah

Client:
Network Dynamics & Simulation Science Laboratory (NDSSL)
CINET is supported by the NSF under Grant No. 1032677

6 December 2012
Overview

- Motivation
- Objectives
- Methodology
- Implementation
- Challenges
- Milestones
- Future Work
Motivation

● Social systems all around us
  ○ Urban transportation systems
  ○ Communication networks
  ○ Epidemics
  ○ Energy systems, etc.

● Social systems as graphs
  ○ Identify critical elements
  ○ Understand inherent phenomenon
  ○ Predict evolving trends
  ○ Intervene with countermeasures, etc.
Objectives

- Enable (semi-)automated graph construction from blogs
  - Support social media analysis research at NDSSL
    - Rumor spread & opinion formation
    - Especially over political blogs
  - Enhance CINET graph repository
    - CINET - A Cyber-Infrastructure for Network Science
Methodology

Blog 1 → Crawlers → Blog Models → Graph Constructor → User Input → Graph(s)

Blog 2

Blog 3

Blog 4

......

Meta Model
Methodology (..contd)

- Start with a seed
- Extract URLs
- Download pages
- Extract content
- Dump content into database
  - As per the metamodel
- Query database
- Construct graphs
Implementation

Data sets

- **Stack Overflow**
  - Most used forum of Stack Exchange family
  - Discusses software programming
  - 3.3M questions
  - 6.6M answers
  - 13M comments
  - 30K tags

- **CNN Political Ticker**
  - Discusses politics
  - Comments are spread across multiple HTML pages

Development setup

- **CPU**
  - Intel(R) Core(TM) i5-2400 CPU @ 3.10GHz

- **Physical Memory**
  - 8GB

- **Disk**
  - 320 GB
  - RPM = 7200
  - Buffer size = 16MB
Graphs Constructed

- **User-user interaction graphs**
  - 1st degree interactions
    - Stackoverflow: post-comment, post-answer and answer-comment
    - CNN political ticker: post-comment
  - 2nd degree interactions
    - CNN political ticker: comment-post-comment

- **Hierarchical representation of blog posts and comments**
Graph Construction: Pseudo Code

- For each blog $B_i$, retrieve all users
- For each user $U_{ij}$, retrieve all posts written by $U_{ij}$
- For each post $P_{ik-j}$, retrieve all comments
- For each comment $C_{it-k}$, get its owner $U_{id}$
- Construct an edge between $U_{ij}$ & $U_{id}$
- For each comment $C_{is}$ written by $U_{ij}$, get its parent post $P_{k'}$
- Construct an edge between $U_{ij}$ & owner of $P_{ik'}$
- Dump the $U_{ij}$ edge list in a file (adjacency matrix)
Stackoverflow
Stack Overflow Approach

- Question
  - Comment
  - Answer
  - Comment

- Post
  - Comment
  - Comment
  - Comment
Hierarchical Representation
Stackoverflow Graphs

Complete user-user interaction graph
- Nodes = 805652
- Edges = 9.3 million
- Max. deg. = 21869
- Avg. deg. = 23.118

Runtime
- Graph construction
  - Tg = 2104 seconds (incl. dumping into file) [7 threads]

Reduced user-user interaction graph
- Nodes = 34265
- Edges = 1 million
- Max. deg. = 5601
- Avg. deg. = 59.5053

Runtime
- Graph construction
  - Tg = 1803 seconds (incl. dumping into file) [1 thread]
Stack Overflow Graph (~35K nodes)

- Nodes = 34265
- Edges = 1 million
- Min. degree = 1
- Max. degree = 5601
- Avg. degree = 59.5

Colors represent modularity classes
CNN Political Ticker
CNN Political Ticker Approach
CNN Constructed Graphs

1st degree interaction graph

- Post-comments
- 270 posts
- Nodes = 2945
- Edges = 5117
- Max. deg. = 943
- Avg. deg. = 3.47

Runtime

- Extraction time
  - $Te = 3120$ seconds (incl. data download)
- Graph construction
  - $Tg = 2$ seconds (incl. dumping into file)

2nd degree interaction graph

- Post-comment and comment-post-comment
- 270 posts
- Nodes = 2945
- Edges = 433630
- Max. deg. = 2120
- Avg. deg. = 294.48

Runtime

- Extraction time:
  - $Te = 3120$ seconds (incl. data download)
- Graph construction
  - $Tg = 6$ seconds (incl. dumping into file)
CNN - First Degree Interaction

- Nodes = 2945
- Edges = 5117
- Min. degree = 1
- Max. degree = 943
- Avg. degree = 3.5

Colors represent modularity classes
Layout used: OpenOrd
Welcome to the Computational Network Sciences (CINET) GRANITE system. [Register here.]

### Input

<table>
<thead>
<tr>
<th>Start Time</th>
<th>Graph</th>
<th>Status</th>
<th>Percent Complete</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dec 4 15:05:31</td>
<td>stackoverflow.gph</td>
<td>Completed</td>
<td></td>
</tr>
<tr>
<td>Dec 4 15:06:31</td>
<td>CNN270_With_Second_Degree.gph</td>
<td>Completed</td>
<td></td>
</tr>
<tr>
<td>Dec 4 15:06:31</td>
<td>ReducedStackoverflow.gph</td>
<td>Completed</td>
<td></td>
</tr>
<tr>
<td>Dec 4 15:08:31</td>
<td>CNN270.gph</td>
<td>Completed</td>
<td></td>
</tr>
</tbody>
</table>

### Results

Results ready for download. Click 'Download Results'.

#### Download Results

\[ |v| = 805652, |E| = 9.31253e+06, \min \text{deg} = 1, \max \text{deg} = 21869, \text{avg deg} = 23.118 \]

*** by Galib 1.0 ***
Challenges

• Extracting content from blogs
  ○ No standardized structure
    ■ Custom crawlers & metamodel based extraction
  ○ Comments are incorporated in different forms
    ■ Only available if signed-in
    ■ Embedded Ajax scripts
    ■ Embedded in the same page
    ■ Focus on embedded comments

• Input flexibility
  ○ Many different ways to specify input
  ○ 1st and 2nd degree interactions to start off with

• Duplicate and anonymous names
  ○ Represented all the anonymous users as one user
Milestones

- Literature Survey - All - 09/25
- Meta Model Creation - HM - 10/03
- Custom Crawlers - HM - 10/31
- Graph Construction - RK - 11/05
- Visualization - RS - 11/20
- Integration & Testing - All - 11/23
- First Demo to Client - All - 11/26
- Fixes & Improvements - All - 12/06
- Final Demo to Client - All - 12/07
- Report Writing - All - 12/10

HM: Hemanth Makkapati
RK: Rushi Kaw
RS: Rajesh Subbiah
Future work

- More blogs
  - different topics
  - different structures
- Generic content extraction
- Richer graph representation
Acknowledgments

- Prof. Fox
  - Help with methodology
- Spencer Lee
  - Graph upload to CINET
- Team CINETGraphViz
  - Graph visualizations
- Keith Bisset
  - Suggestions on data sets
References


Thank you!