SOCIAL-MEDIA NETWORK COLLECTION PROBLEMS

Web Crawling
Semantic Graph Sampling

Jeremy D. Wendt
with Rich Field, Daniel Garcia, Tu-Thach Quach, Randy Wells, David Zage, Cindy Phillips, Ali Pinar, Sucheta Soundarajan
WHY WEB CRAWLING?

• Isn’t the Internet so 2000?

• Still a significant part of search result ranking

• Situational awareness
  • Although social media are hot, traditional websites and blogs still contain much relevant data
  • On a “deliberative” scale, micro-blogging < blogging < editorials/traditional news

• I’d really love the text and graph of topically relevant pages
queue ← seed sites

while (!queue.isEmpty()) {
    download and parse next page
    store relevant data //text, metadata, images...
    if (continuation criteria met)
        queue ← all links found on page
}

perform analysis // e.g., [Colbaugh/Glass, 2010]
PROBLEMS

• Search Engine Optimization (SEO) causes websites to be designed to trap crawlers in a domain

• No way to determine relevance of a linked page without retrieving it

• Rapid explosion in links limits depth crawl can reach in reasonable time

• Many graph analytics become less useful
  • Instead of showing communities of same-topic pages, community detection finds domains (sometimes merging highly-related domains)
Example 1

• ~240 links on page
• 4 links within article
• 17 links appear to sources outside WP domain
• Only 5 appear to link to other movie reviews or relevant content

Example 2

- ~190 links on page
- 0 links within article
- 27 links appear to sources outside fatcyclist domain
- None appear to link to other movie reviews or relevant content
• Google-driven crawls
  • Crawl only sites returned by Google search plus one hop out?
    • Loses the connected graph that you get from standard crawls

• Topic-based continuation criteria
  • Halting branches of crawls if current page doesn’t match crawl criteria term list
    • Requires an extensive list of “approved” words, will have both false positives and false negatives
• **Gold Standard**
  - If hosting a domain, you know all links that were followed to arrive at any page on your site
  - Most-often followed links are more “real” than links than those never followed

• **Dueling Community Detection**
  - Run comm. det. on the graph, LDA on the text
  - Find analytics that reweight the graph edges to get Louvain and LDA to converge

• **Analyze graph and HTML properties of edges/links**
• Twitter represents a very interesting social graph
  • All nodes are users (not necessarily people)
  • Multiple edge types: friends, followers, @refers, retweets, etc.
  • Information flows through this graph at different rates depending on link type, complex “physics”, etc.

• If we can understand this space, we can better understand what memes will catch and take off
  • Percolation and information spread
Twitter API

- Twitter allows anyone free access to their data
  - Severely rate limited
  - Different rates for different query types
- 305M active users*
  - >580 years to get all of those

Graph sampling in traditional graphs well explored
  - e.g., [Leskovec, Faloutsos 2006]; [Maiya, Berger-Wolf 2011] describe tradeoffs for different sampling techniques

Two critical parts for graph sampling
  - What measures do you wish to preserve?
  - Which techniques can you use to sample?

Semantic graphs add new features
  - New measures available
  - Each edge type need not be sampled using same technique
• Undirected, single edge type
  • Choose a per-node measure $D$ (e.g., degree)
  • $E\{D\}$ – average degree, $P[D < a]$ – distributions

• Undirected, multiple edge type
  • Choose a per-node measure for each edge type $i$: $D_i$
  • $E\{D_i\}$ and $P[D_i < a]$ are now marginals
  • Joint measures now available: correlation($D_i$, $D_j$)
TRADITIONAL SAMPLING

Node visited

Node discovered, but not visited
SEMANTIC SAMPLING

- Node visited by all 3
- Node visited by 2
- Node visited by 1
- Node discovered, but not visited
• Web Crawling
  • Topically relevant graph
  • Which graph measures might indicate more relevant page vs. less relevant to topic on current page?
  • We’d be as happy to eliminate less relevant linked pages as to identify most relevant linked pages

• Semantic Graph Sampling
  • Information percolation
  • How do we formalize “visited” with multiple visit types?
  • When are marginal or joint distributions more relevant?
  • Mixed sampling strategies for different edge types…

Partial funding of this work provided by Sandia Laboratory-Directed Research & Development funding.