Analyzing the Semantic Modeling Capabilities of Google Sets

Azfar Khandoker
Ryan Miller

April 23, 2013

Background

- Google Sets
  - Google Labs (2002)
  - Semantics behind search queries
  - Discontinued in 2011
  - Lives on in Google Spreadsheets!

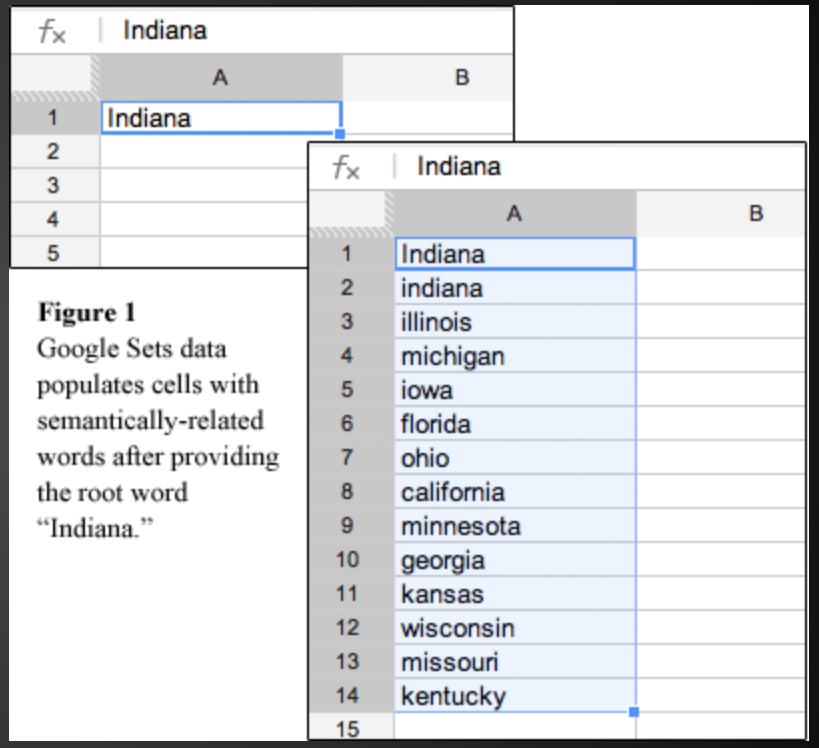


Figure 1
Google Sets data populates cells with semantically-related words after providing the root word "Indiana."

Background

- Building a star graph using Google sets
  - Start with seed word, *s*
  - Use Google Sets to retrieve result set $R = \{ r_1 , \ldots , r_n \}$

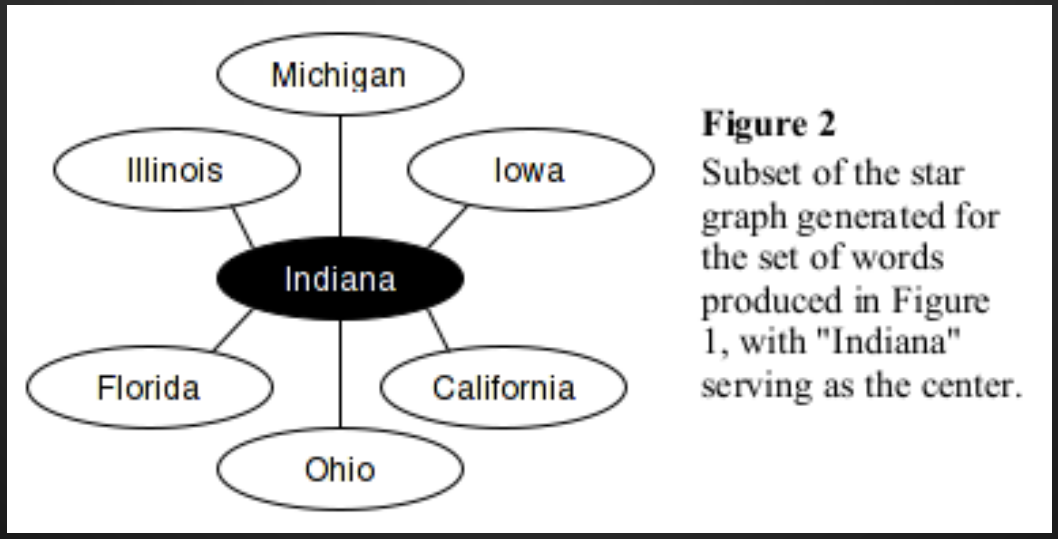


Figure 2
Subset of the star graph generated for the set of words produced in Figure 1, with "Indiana" serving as the center.

Background

- Other Semantic Networks
  - Wikipedia [1]
    - 5.7M nodes, 130M edges
    - Avg. Degree: 45.531 (max_{IN} = 374934 = "U.S.")
    - Avg. Path Len: ???
  - Amazon Co-purchasing [2]
    - 542,000 nodes, 1.2M edges
    - Avg. Degree: 4.538 (max = 118 = "Laura")
    - Avg. Path Len: 2.842

[1] Haselgrove, H. "Using the Wikipedia page-to-page link database." Retrieved from http://haselgrove.id.au/wikipedia.htm.

[2] J. Leskovec, L. Adamic and B. Adamic. The Dynamics of Viral Marketing. ACM Transactions on the Web (ACM TWEB), 1(1), 2007. Retrieved from http://snap.stanford.edu/data/amazon-meta.html

Background

- Other Semantic Model Research
  - Small World Phenomenon (Milgram, 1967; Watts & Strogatz, 1998)
    - Small average path length, high clustering
    - Power law degree distribution

$$p(x) \propto x^{-\alpha}$$

  - Latent Semantic Analysis (Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998)
    - Cosine similarity

$$\sigma_{ij} = \cos\theta = \frac{n_{ij}}{\sqrt{d_i d_j}}$$

$$n_{ij} = \sum_k A_{ik} A_{kj}$$

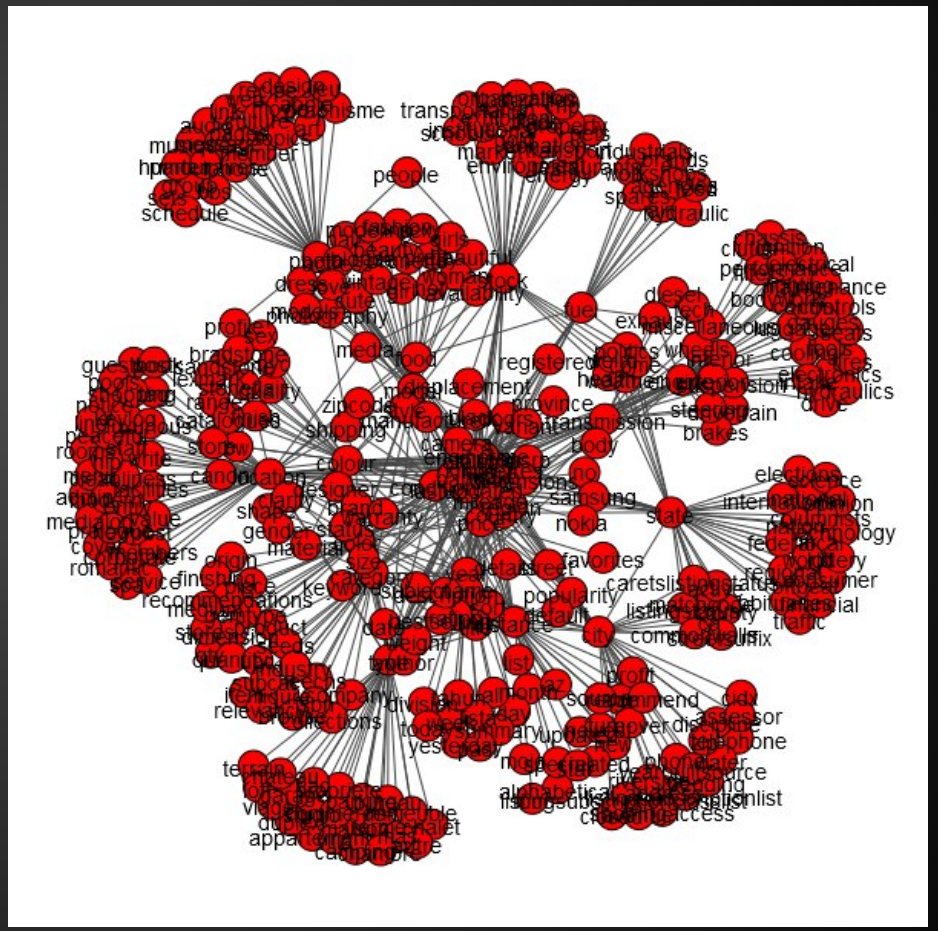
Implementation

- Automate Google Sets output
- Build network starting from a *seed word*

- When to stop?
  - Two ideas implemented:
    - *Destination word* (perfect to relate two seemingly-unrelated words)
    - *Depth level* (better for general network analysis)

Implementation

- Building a network using Google sets
  - With repeated iterations, we generate a graph
  - Nodes represent words
  - Edge (*i*, *j*) means $word_j$ was a result of seed $word_i$



Graph of "mpg" to "oil"

Studies

- Shortest Path Characteristics
- Comparison of Semantic Networks
- Semantic Similarity

Studies

- Shortest Path: "wine" --> "france"

  Paths found:

Network	**Path Length**	Path
Google Sets	4	"wine" --> "champagne" --> "bordeaux" --> "france"
Wiki	1	"Wine" --> "France"
Amazon	4	"Wine" --> "Italy (Culinaria)" --> "Culinaria: The United States: A Culinary Discovery (Culinaria)" --> "Culinaria France (Culinaria Series)"

Studies

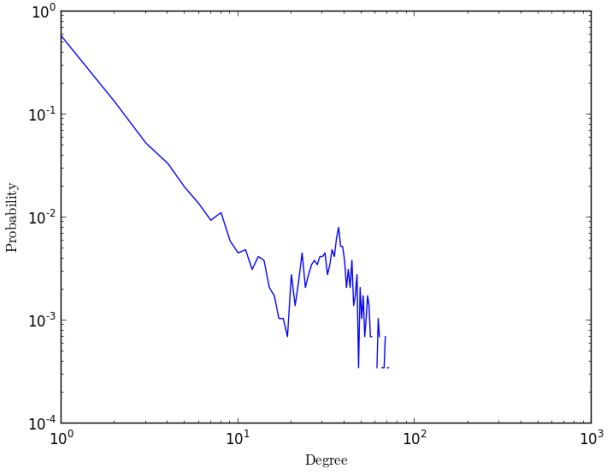
- Shortest Path Characteristics
  - Small average path length, high clustering coefficient

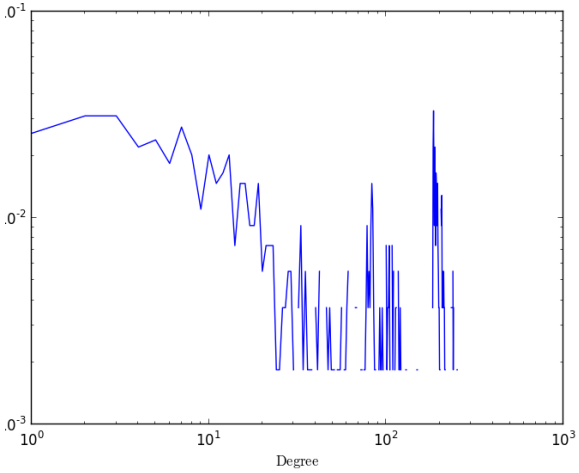
	Google Sets	Wiki	Amazon
# nodes	2,871	544	1,373
# edges	8,962	25,403	3,432
Avg. degree	6.243	93.393	5.998
Avg. path length	4.329	2.344	4.287
Clustering coeff.	0.255	0.727	0.343

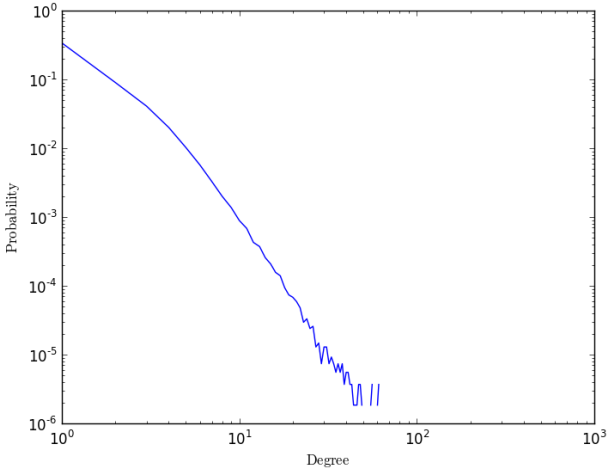
Studies

- Small World Characteristics
  - Power law degree distribution

Google Sets ✓ Wiki ✗ Amazon ✓



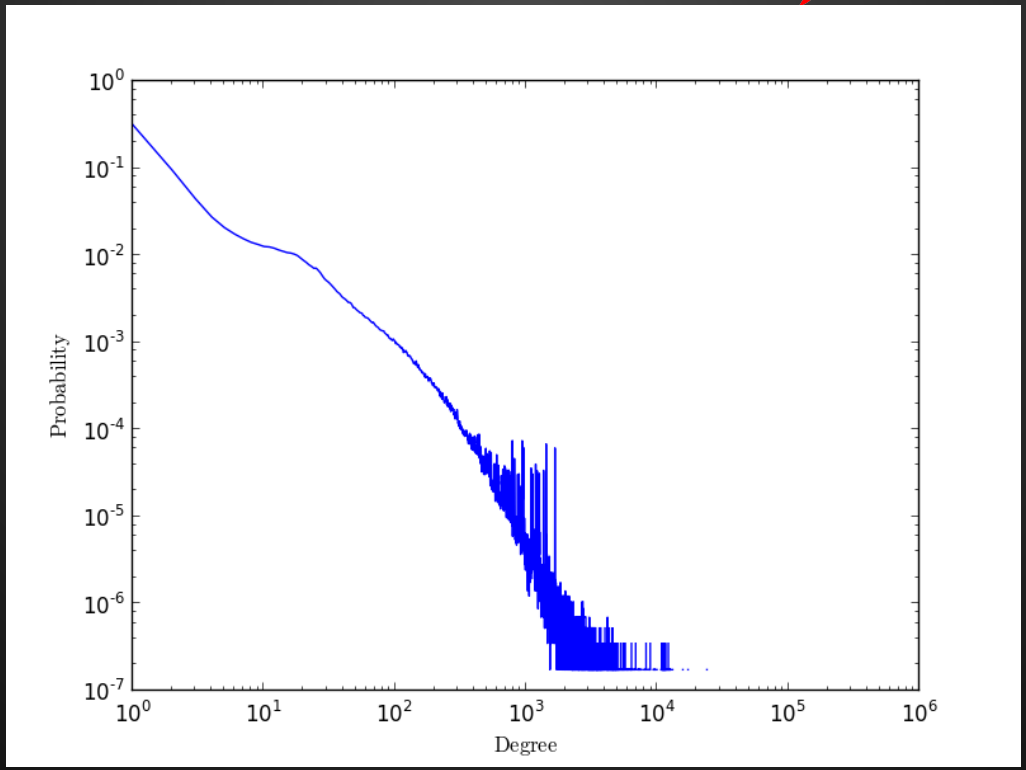




Studies

- Shortest Path Characteristics
  - Power law degree distribution

Entire Wiki Network



Studies

- Comparison of Semantic Networks - "Oreo"

	Google Sets	Wiki	Amazon
# nodes	10,229	16,478	1,119
# edges	33,081	1,252,227	1,118

	Unique for Google Sets	Not in Google Sets	Shared	Percentage Shared
Google Sets vs. Wiki	9,682	15,931	547	3.320 %
Google Sets vs. Amazon	9,919	809	310	27.703 %

Studies

- Semantic Similarity
  - Cosine similarity

$$\sigma_{ij} = \cos\theta = \frac{n_{ij}}{\sqrt{d_i d_j}}$$

$$n_{ij} = \sum_k A_{ik} A_{kj}$$

	Google Sets	Wiki	Amazon
wine -> france	0.0	0.0058	0.0096
wine -> grape	0.0	0.02229	0.0344
wine -> alcohol	0.1427	0.00921	0.0094
wine -> beer	0.8007	0.1987	0.6862

Conclusions

- Findings
  - Google Sets and Amazon both exhibit small world behavior
  - Overlap between semantic models for Google Sets and Amazon implies a use of Google Sets data for product recommendation
  - Cosine similarity can further refine recommendations by semantic meanings

Conclusions

- Future Work
  - Compare against other networks
    - Twitter
    - Facebook
    - The New York Times
  - Psychology
    - Word association of Humans vs. Google Sets

Questions?