Semantic Entity Resolution

Jeremiah Greer

Outline

- CERCIS
- Entity Resolution
- Semantic Contribution
- Google Search
- Neo4j Graph
- Truth Set Survey
- Additional Features/Future Work

CERCIS

Cincinnati Entity Resolution, Combination, and Information System

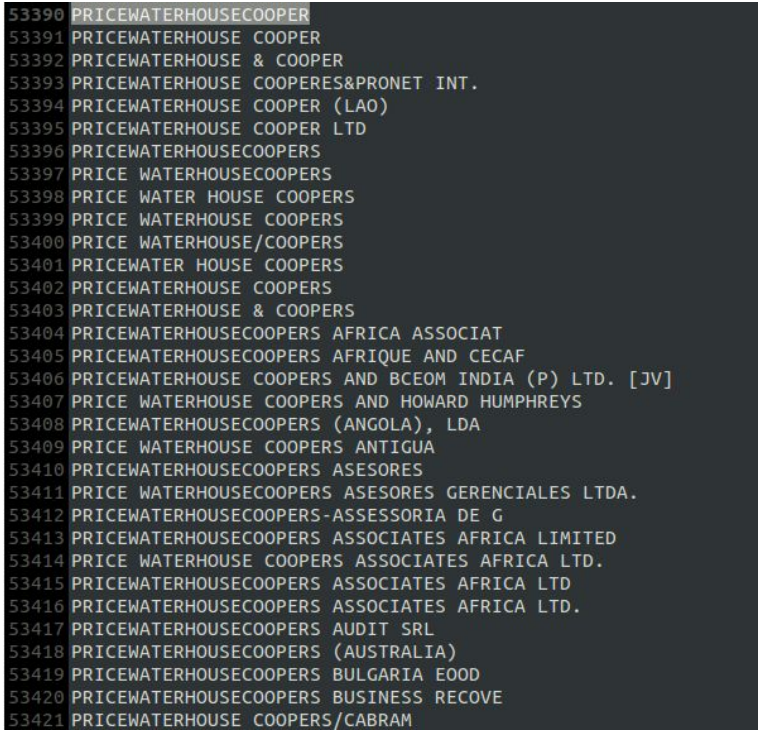
Identifying Companies

Resolving into single entities

Discrepancies/Fraud Detection

CERCIS

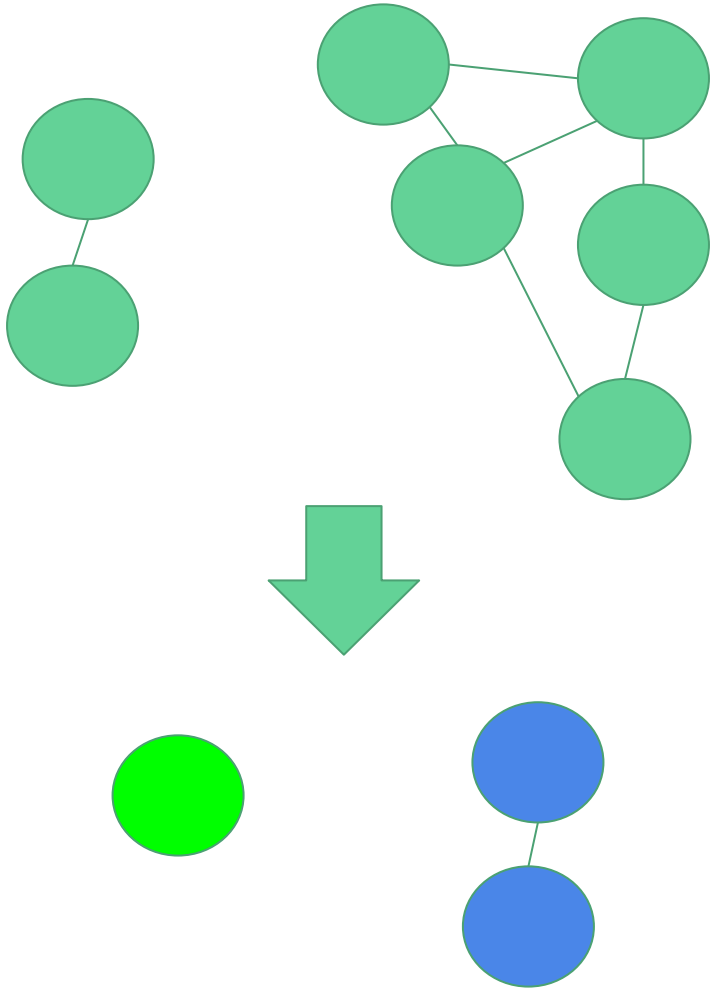
- World Bank “Major Contract Awards” Dataset
  - No address or distinct identifiers
  - Missing cells
  - Only consistent point is Supplier Name
- Company names are inconsistent
  - Input by individuals
  - Different representations of the same name
- Prevent giving loans to fraudulent companies
  - Purposeful misspelling of names
  - Name changes
  - Predict fraudulent behavior



Entity Resolution

Resolving Entities

- Company names are the consistent minimum
- Syntactic and Semantic Resolution
- Build relationships between entities
- Resolve entities to their single identities
- Results in unique, non-duplicate entities for reference



Semantic Contribution

Syntactic concerns the physical structure of the name

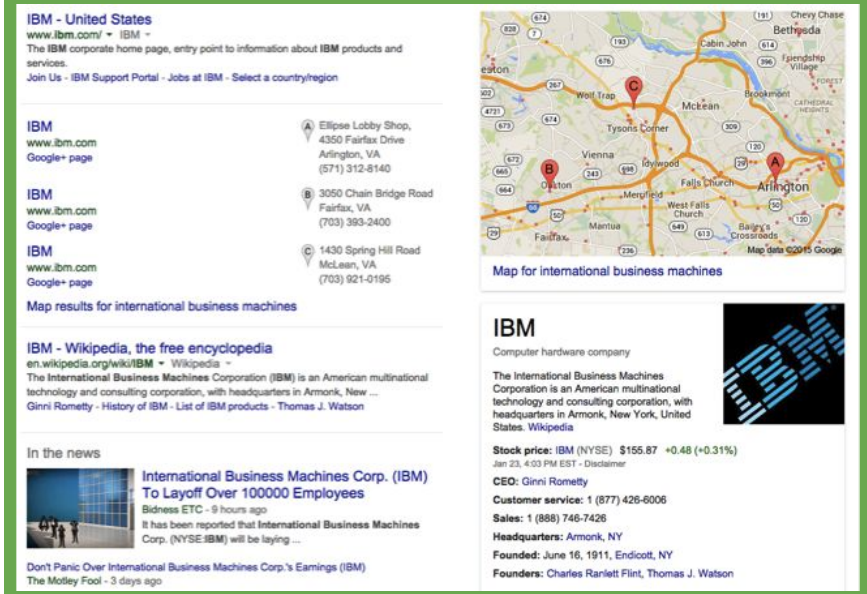
Semantic attempts to find what the name means/represents intuitively

Semantic Contribution

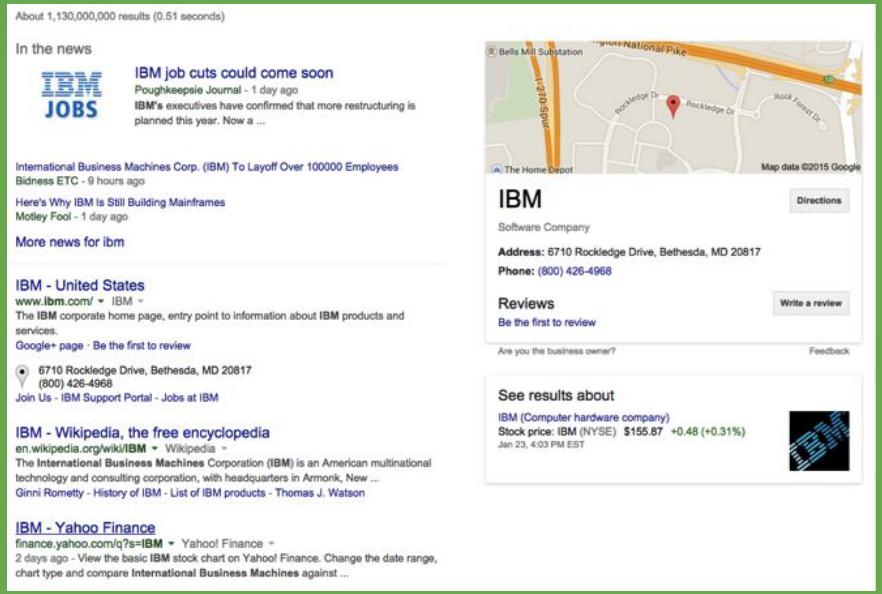
- Syntactic attempts to find similarity between names, but sometimes that can fail
  - Complete name changes: Valujet, Airtran => Southwest
  - Alternate representations/abbreviations: SMEC, Snowy Mountain Engineering Corporation
  - Individual people’s names
- Semantic finds missing connections by meaning of terms
  - IBM and International Business Machines
- Variety of methods/features available

Semantic Contribution

International Business Machines

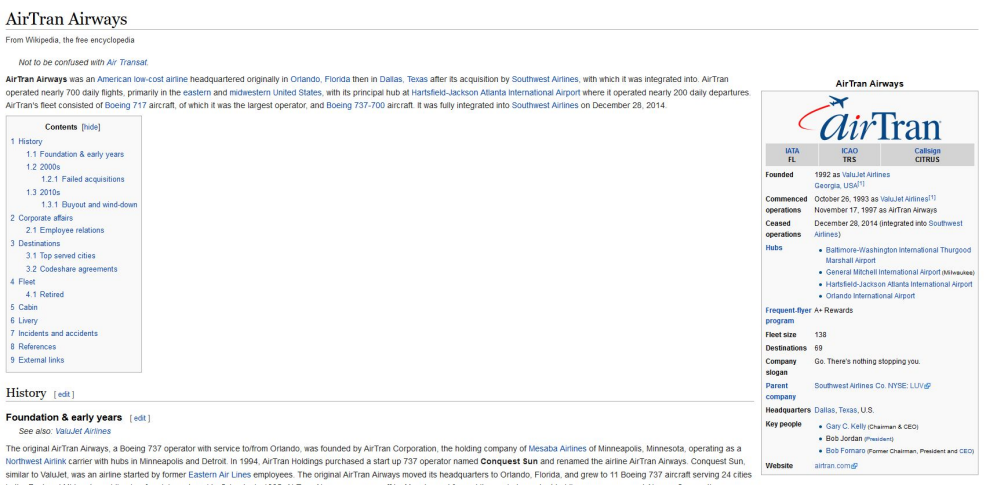


IBM



Semantic Contribution

- Valujet, AirTran, Southwest
- No Google Search Connection
- No Syntactic Connection
- Wikipedia Connection



AirTran Airways

From Wikipedia, the free encyclopedia

Not to be confused with Air Transat.

AirTran Airways was an American low-cost airline headquartered originally in Orlando, Florida then in Dallas, Texas after its acquisition by Southwest Airlines, with which it was integrated into. AirTran operated nearly 700 daily flights, primarily in the eastern and midwestern United States, with its principal hub at Hartsfield-Jackson Atlanta International Airport where it operated nearly 200 daily departures. AirTran's fleet consisted of Boeing 717 aircraft, of which it was the largest operator, and Boeing 737-700 aircraft. It was fully integrated into Southwest Airlines on December 28, 2014.

Contents [hide]

1 History
- 1.1 Foundation & early years
- 1.2 2000s
  - 1.2.1 Failed acquisitions
- 1.3 2010s
  - 1.3.1 Buyout and wind-down

2 Corporate affairs
- 2.1 Employee relations

3 Destinations
- 3.1 Top served cities
- 3.2 Codeshare agreements

4 Fleet
- 4.1 Retired

5 Cabin
6 Livery
7 Incidents and accidents
8 References
9 External links

History [edit]

Foundation & early years [edit]

See also: ValuJet Airlines

The original AirTran Airways, a Boeing 737 operator with service to/from Orlando, was founded by AirTran Corporation, the holding company of Mesaba Airlines of Minneapolis, Minnesota, operating as a Northwest Airlink carrier with hubs in Minneapolis and Detroit. In 1994, AirTran Holdings purchased a start up 737 operator named **Conquest Sun** and renamed the airline AirTran Airways. Conquest Sun, similar to ValuJet, was an airline started by former Eastern Air Lines employees. The original AirTran Airways moved its headquarters to Orlando, Florida, and grew to 11 Boeing 737 aircraft serving 24 cities

AirTran Airways

IATA FL	ICAO TRS	Callsign CITRUS
Founded	1992 as ValuJet Airlines Georgia, USA[1]
Commenced operations	October 26, 1993 as ValuJet Airlines[1] November 17, 1997 as AirTran Airways
Ceased operations	December 28, 2014 (integrated into Southwest Airlines)
Hubs	• Baltimore-Washington International Thurgood Marshall Airport • General Mitchell International Airport (Milwaukee) • Hartsfield-Jackson Atlanta International Airport • Orlando International Airport
Frequent-flyer program	A+ Rewards
Fleet size	138
Destinations	69
Company slogan	Go. There's nothing stopping you.
Parent company	Southwest Airlines Co. NYSE: LUV
Headquarters	Dallas, Texas, U.S.
Key people	• Gary C. Kelly (Chairman & CEO) • Bob Jordan (President) • Bob Fornaro (Former Chairman, President and CEO)
Website	airtran.com

ValuJet Airlines

From Wikipedia, the free encyclopedia

This article **needs additional citations for verification**. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed. *(November 2007)*

ValuJet Airlines was an American low-cost carrier, headquartered in unincorporated Clayton County, Georgia,[2] that operated regularly scheduled domestic and international flights in the Eastern United States and Canada[3] during the 1990s. The company was founded in 1992 and was known for its cost-cutting measures. All of the airline's planes were purchased used from other airlines, very little training was provided to workers, and contractors were used for maintenance and other services. The company quickly developed a reputation for its lax safety. In 1995, the military refused ValuJet's bid to fly military personnel over safety worries, and officials at the FAA wanted the airline to be grounded.

The May 1996 crash of Flight 592, caused by illegally stored hazardous materials on board, spelled doom for the airline. ValuJet was grounded the next month and not allowed to fly again until September, with a greatly reduced fleet. The airline's major customers never returned and the company suffered major losses. In 1997, the company merged with the much smaller regional airline AirWays Corp., holding company for AirTran Airways. Although ValuJet was the nominal survivor, the merged airline operated as AirTran until its merger with Southwest Airlines in 2011. AirTran never acknowledged its past involvement with ValuJet and kept the airline's memorabilia locked in a warehouse in Atlanta.

Contents [hide]

1 History
- 1.1 Inception
- 1.2 Safety problems
- 1.3 Fallout from the crash of Flight 592

2 Fleet
3 Destinations
4 Incidents and accidents
- 4.1 Flight 597
- 4.2 Flight 592

5 References
6 External links

History [edit]

Inception [edit]

ValuJet was founded in 1992 and began operations on October 26, 1993. It originally offered service from Atlanta to Orlando, Jacksonville and Tampa with a single Douglas DC-9 that previously belonged to Delta Air Lines. The first flight, Flight 901, flew from Atlanta to Tampa. The carrier was headed by a group of industry veterans including co-founder and chairman Robert Priddy, who had started a string of successful airlines including Atlantic Southeast Airlines (ASA), Air Midwest Airlines, and Florida Gulf Airlines. Board members Maury Gallagher and Tim Flynn, the other co-founders, developed and ran

ValuJet Airlines

IATA J7	ICAO VJA	Callsign CRITTER
Founded	1992 Georgia, USA[1]
Commenced operations	October 26, 1993[1]
Ceased operations	November 17, 1997[1] (AirTran Airways)
Hubs	• Hartsfield-Jackson Atlanta International Airport • Logan International Airport (Boston) • Miami International Airport • Orlando International Airport • Philadelphia International Airport • Washington Dulles International Airport
Fleet size	56
Destinations	28
Headquarters	Clayton County, Georgia, USA
Key people	Robert Priddy (Chairman) Maurice J. Gallagher, Jr. (CEO & President)
Website	valujet.com

Google Search

Basis

- Multiple representations (names) of same entity
- Google provides results “most relevant” to user
- Page rank, search selection, etc.
- Searching IBM and International Business Machines returns largely same results
- First results are considered most relevant

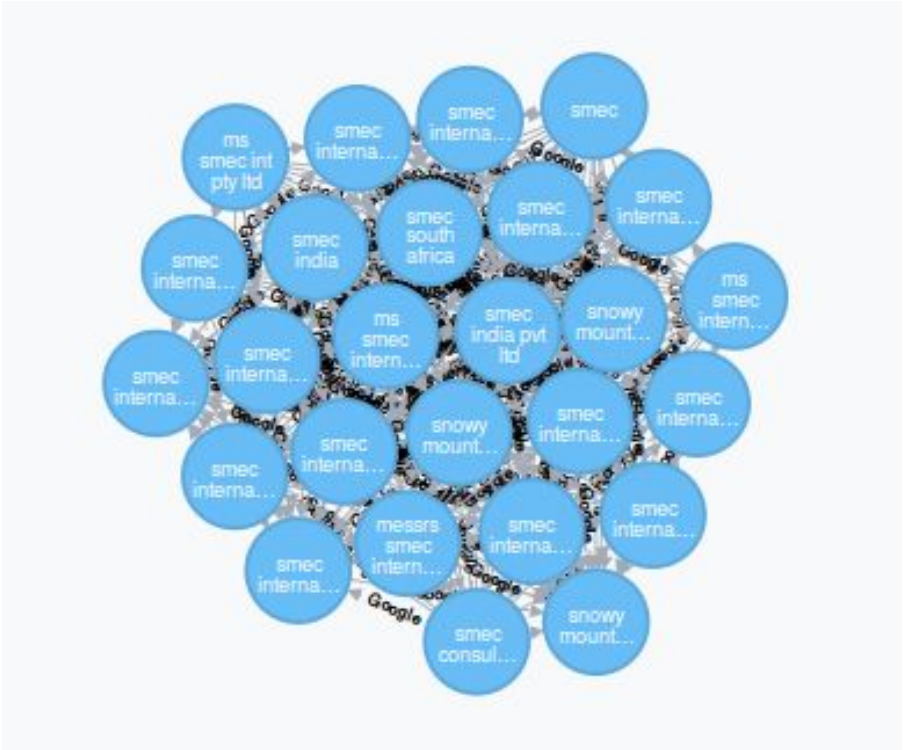
Implementation

1. Entity names are normalized and duplicates are removed
2. Search using normalized name on Google, scrape first 10 result links
   a. Want to keep only the most pertinent links, few go beyond 1st page
3. Store result links, get number shared between 2 entities (intersection)
4. Add connection to Neo4j Graph
5. Query

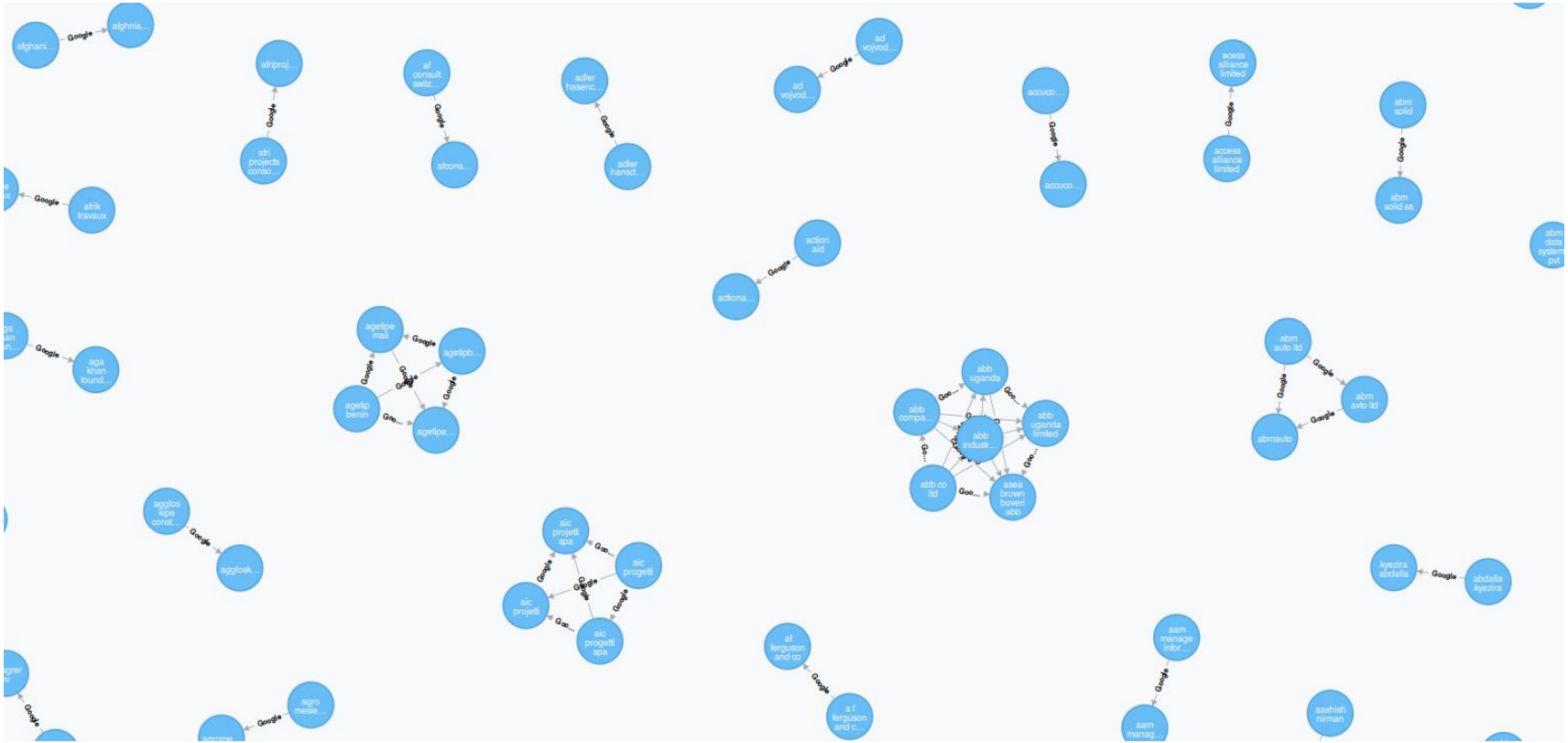
14th engineering bureau of china railway	5	china railway 16th bureau group 1st engineering coltd
14th engineering bureau of china railway	4	china railway 13th engineering bureau group co ltd
14th engineering bureau of china railway	4	china railway 20th bureau co ltd
14th engineering bureau of china railway	4	china railway 4th bureau group corporation
14th engineering bureau of china railway	4	china railway bureau group co ltd
14th engineering bureau of china railway	4	china railway bureau no 18 cor ltd group
14th engineering bureau of china railway	4	china railway construction 14 bureau
14th engineering bureau of china railway	4	china railway sixteen bureau group co
14th engineering bureau of china railway	4	the 17th engineering bureau of china railway the 1st enginee
14th engineering bureau of china railway	3	china railway 10th bureau group coltd
14th engineering bureau of china railway	3	china railway 12th bureau group co ltd
14th engineering bureau of china railway	3	china railway 15 bureeau group coporation
14th engineering bureau of china railway	3	china railway 19th bureau group corp
14th engineering bureau of china railway	3	china railway 19th bureau group corp ltd
14th engineering bureau of china railway	3	china railway 20th bureau group co ltd
14th engineering bureau of china railway	3	china railway 25th engineering bureau group coltd
14th engineering bureau of china railway	3	china railway bureau for large bridge corp ltd group

Neo4j Graph

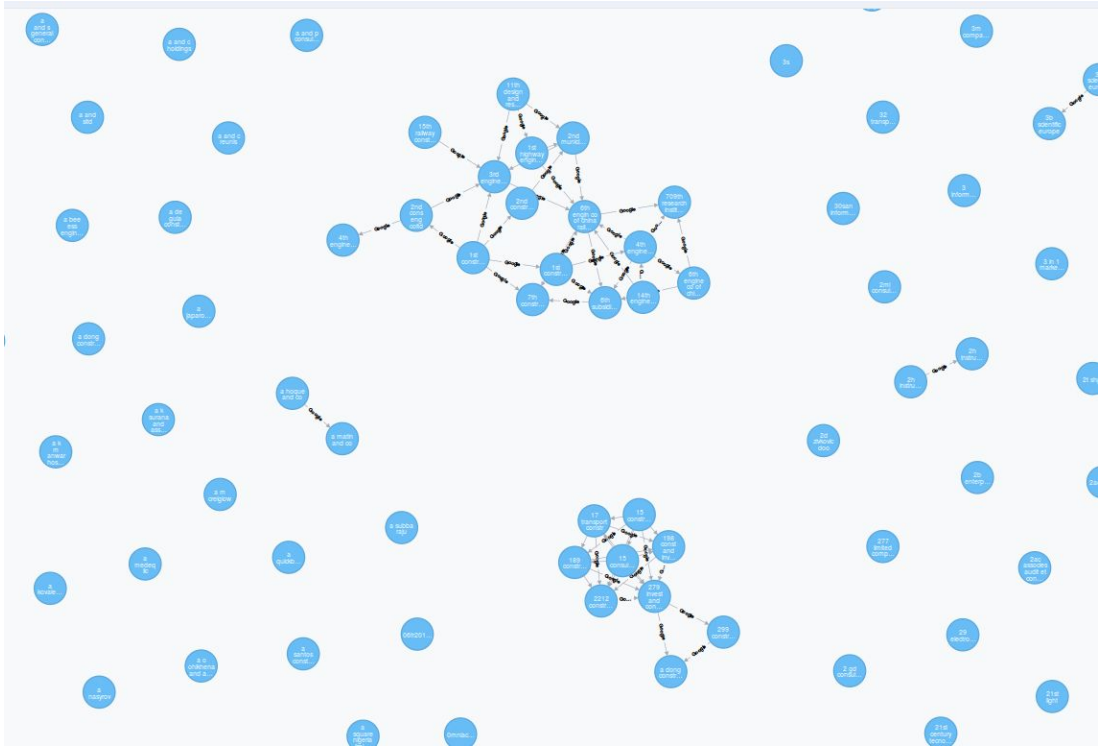
SMEC, r.score >= 2, limit 100



All, r.score >= 9, limit 100



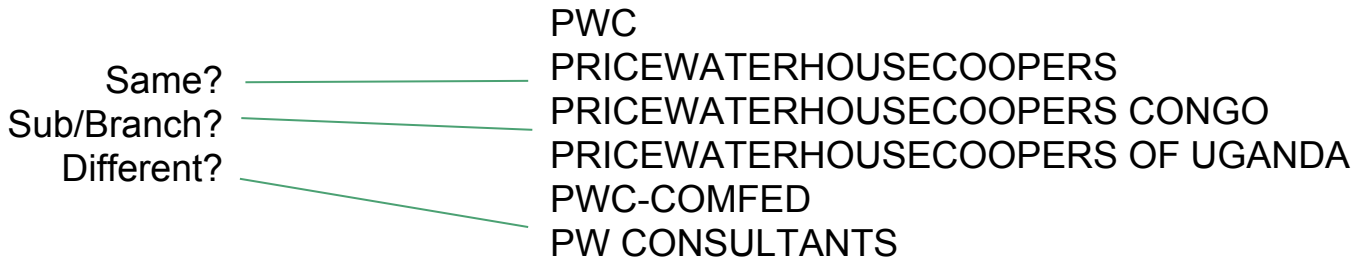
General Graph, limit 200



Truth Set/Survey

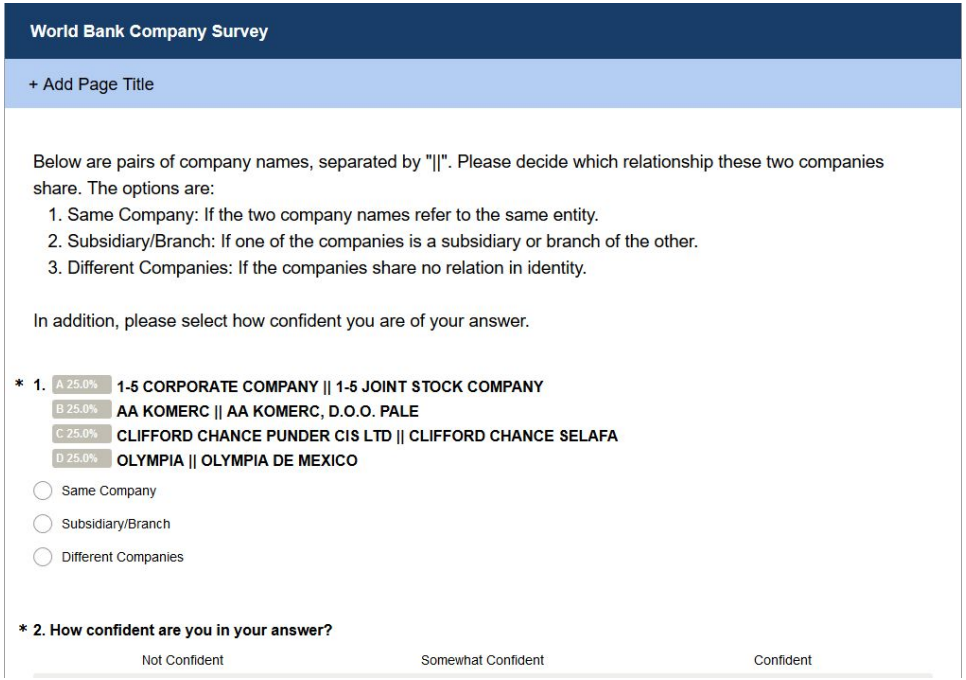
Problem Areas

- Currently have no knowledge of “correct” clusterings
- Companies believed to have three possible relationships:
  - Same
  - Subsidiary/Branch
  - Different
- Need a “Truth Set” to determine accuracy



Survey

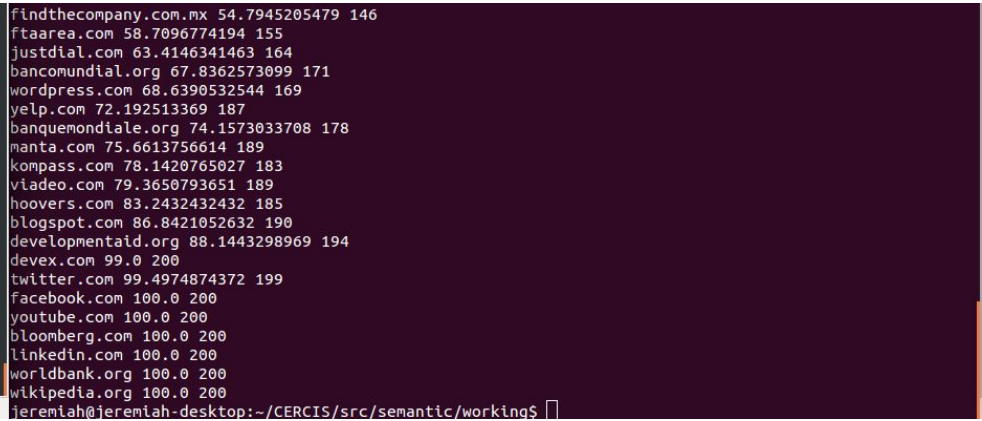
- Relationship between two entities
- Confidence level
- Mixture of random and chosen
- Currently using SurveyMonkey
- Dynamic survey development



Additional Features

Top-Level Domains

- Each of the 10 urls saved per entity has a TLD
- Potential additional identifier (ibm.com, pwc.com, etc.)
- Problem:
  - Important (ibm.com)
  - Non-Important (wikipedia.org)



Wikipedia Scrape

- Find Wikipedia page
- Scrape links and terms
- Build connections
- Problems:
  - Wikipedia Search
  - Wikipedia Urls