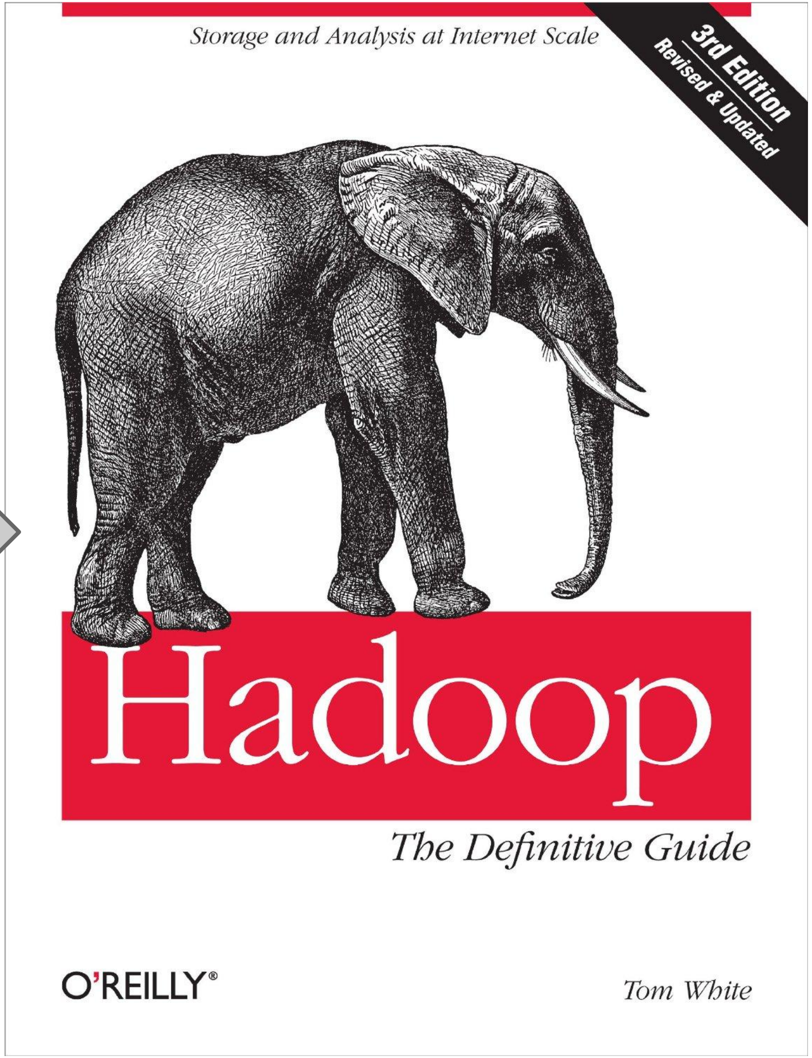
Yelp progress report

Mitsue Iwata, Kyle Magida, Scarlett Swerdlow

Association rules, generally

- Consider two binary variables: x_a and x_b.
- If $x_b = 1$ more often when $x_a = 1$, then $x_a \Rightarrow x_b$ is an association rule.

Building Effective Algorithms and Analytics
for Hadoop and Other Systems
MapReduce Design Patterns
O'REILLY®
Donald Miner & Adam Shook

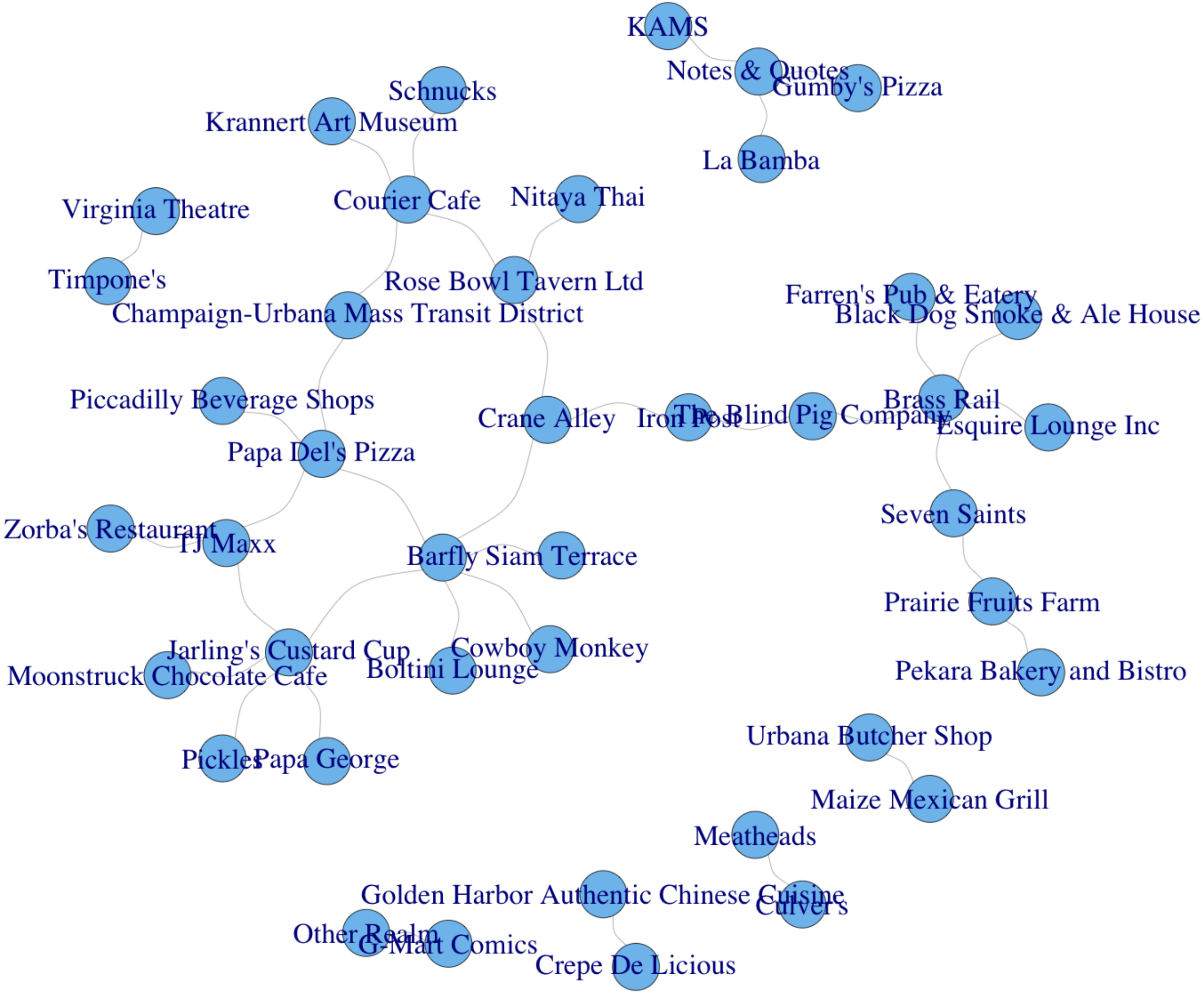

Storage and Analysis at Internet Scale
3rd Edition
Revised & Updated
Hadoop
The Definitive Guide
O'REILLY®
Tom White

Possible Yelp association rules

- People who like Plein Air also like Robust
- People who don’t like Zberry do like Red Mango
- People who like Valois in Chicago also like Tastees Diner in Bethesda

Proof of concept: Champaign-Urbana

- People who go to Culver's are 38 times more likely to go to Meatheads than people who don’t go to Culver’s.


KAMS
Notes & Quotes
Gumby's Pizza
La Bamba
Schnucks
Krannert Art Museum
Courier Cafe
Nitaya Thai
Virginia Theatre
Timpone's
Rose Bowl Tavern Ltd
Champaign-Urbana Mass Transit District
Farren's Pub & Eatery
Black Dog Smoke & Ale House
Piccadilly Beverage Shops
Crane Alley
Iron Post
The Blind Pig Company
Brass Rail
Esquire Lounge Inc
Papa Del's Pizza
Seven Saints
Zorba's Restaurant
TJ Maxx
Barfly
Siam Terrace
Prairie Fruits Farm
Jarling's Custard Cup
Cowboy Monkey
Boltini Lounge
Moonstruck Chocolate Cafe
Pekara Bakery and Bistro
Urbana Butcher Shop
Pickles
Papa George
Maize Mexican Grill
Meatheads
Golden Harbor Authentic Chinese Cuisine
Culver's
Other Realm
G-Mart Comics
Crepe De Licious

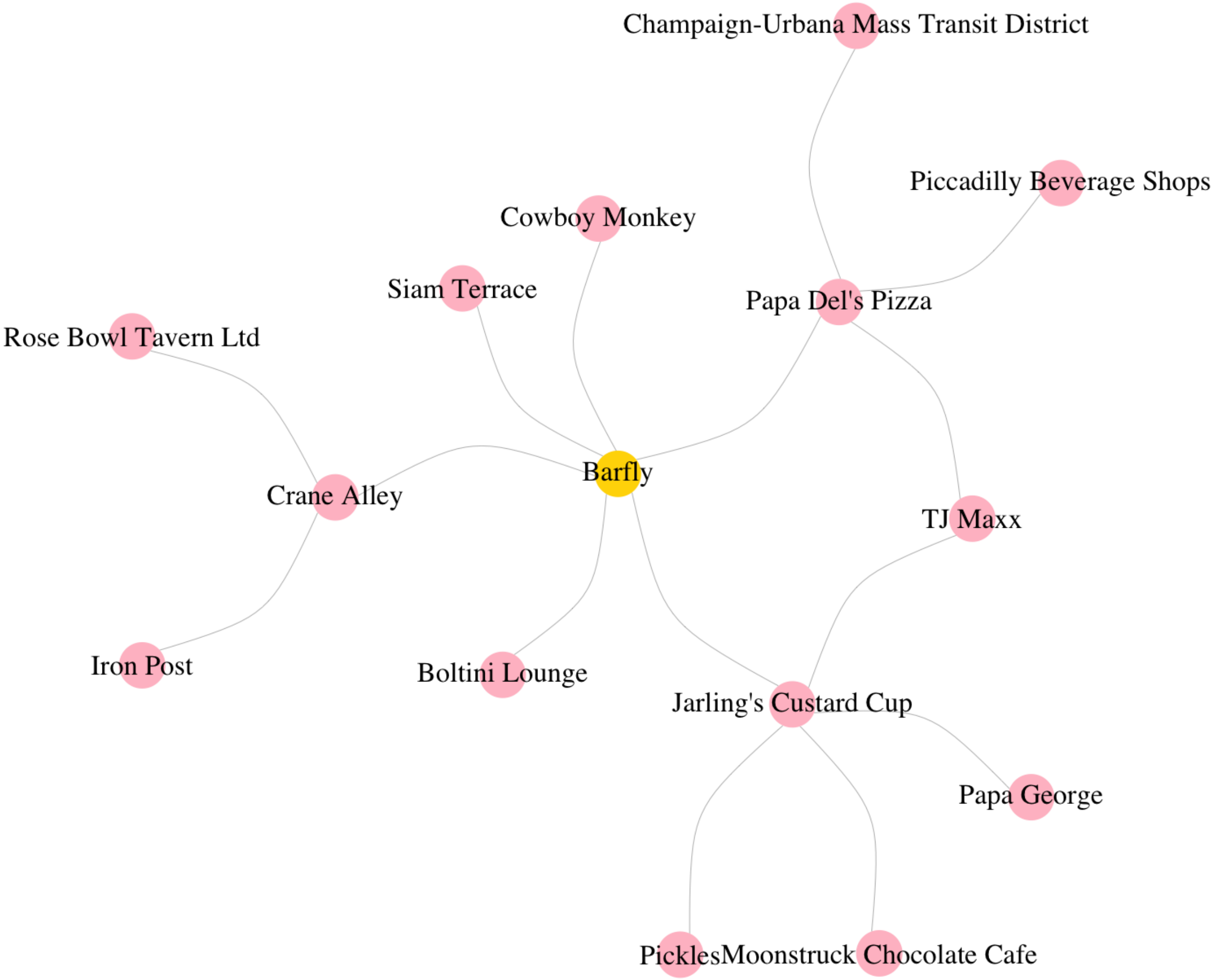
Most “connected” businesses

Degrees:

- Barfly (6)
- Brass Rail (5)
- Jarling’s Custard Cup (5)

Betweenness:

- Crane Alley (193)
- Barfly (187)
- Iron Post (152)


Champaign-Urbana Mass Transit District
Piccadilly Beverage Shops
Cowboy Monkey
Siam Terrace
Papa Del's Pizza
Rose Bowl Tavern Ltd
Crane Alley
Barfly
TJ Maxx
Iron Post
Boltini Lounge
Jarling's Custard Cup
Papa George
PicklesMoonstruck Chocolate Cafe

Scaling up: pairs of businesses

- About 3.7 billion possible pairs of businesses
- Not every pair of businesses will have a shared review
- Sample of 2000 reviews yielded 139 businesses and 2300 pairs

Association rule steps (Map Reduce)

- Support: P(x)
  - Map every review keyed with business & return sum
- Confidence: P(x|y)
  - Create user level data through building lists of all reviews
  - Reduce list to pair key value run
  - Map reduce again to sum
- Lift: Confidence/Support = P(x|y)/P(x)
  - Sum through Map Reduce keyed on each business

Yelp business network analysis

- Node: Business
- Edge: User review
- Use association rules to connect nodes
- Identify most connected businesses
    - degrees - number of edges
    - betweenness - bridge between two nodes
- What do they have in common?