Proposal

Introduction

I'm interested in looking at how companies interact on the internet. I've found a community created dataset about Businesses and the other businesses their websites link to. I'm also looking for another dataset that is a bit cleaner and has information about corporate ownership.

Dataset

Name: Relato Business Graph Database

Link: https://data.world/datasyndrome/relato-business-graph-database

As per the description on the site:

"It contains links between businesses pulled from the web. It contains 373,663 links between companies, of the type's "partnership" (one company listed on another company's partnership page), "customer" (one company listed on another company's example customer page), "competitor" (co-bidders on AdWords above some limit), "investment" (a company listed on a VC's website), "supplier" (the inverse of the "customer" type. This dataset was used to drive both a lead generation system where metrics on the graph fed into a classification for leads (lead/no lead) and a market visualization system (a force directed layout of markets and their segments)."

Analysis

Fundamentally, I want to do cluster and traffic analysis. I want to determine how important certain nodes are, and their importance in controlling the flow of traffic. Knowing these some interesting outcomes can come through. I think I can get a better understanding of company relationships based on their websites. I can find clusters for example based on companies and their subsidiaries and determine how the traffic within those clusters operates. I can get a better sense of industry relationships from the links between companies that are not of the same ownership.

Technologies

Python, Networkx, Spark (via Databricks), potentially Neo4j