Branching Out: Graph Theory Fundamentals

Authors
Eric Oak
857-275-0591
eric.oak@spglobal.com

Chris Rogers
(44) 7526-556-853
christopher.rogers@spglobal.com

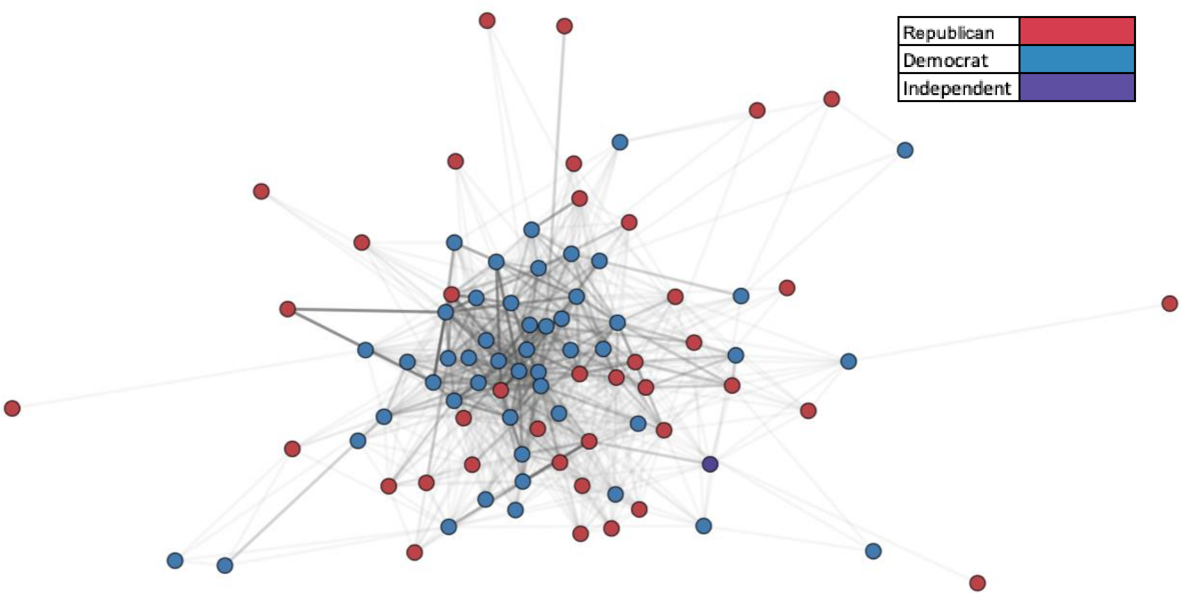
Investment analysis has evolved beyond financial data to non-financial, or alternative data. Typically, the focus has been on using alternative datasets that are purely time-series and tabular, which is suited to stock selection models. Graph networks meanwhile offer investors the ability to gain deeper insights into the connections between economies, industries, and individual corporations. Examples of networks include corporate stock holdings, analysts' stock coverage, management board memberships, political relationships, common patents, and supply chains.

Graph networks give additional insights as:

- Networks can be created out of almost any dataset that captures relationships.
- Network analysis uncovers insights hidden in tabular data: the most influential person in your network, how two companies are connected, structural holes in supply chains, or a key employee.
- Advanced attributes such as betweenness centrality model concepts like importance, identifying key objects. This can help make decisions or find nodes for further analysis such as important suppliers or executives to investigate in more detail.

This report introduces the fundamentals of graph theory by modeling the relationships among members of the 112th U.S. Senate through their sponsorships of finance bills that were passed (Figure1). The visualization shows Senators with more connections (influence, importance) in the center while Senators with fewer connections are on the outside. This shows the influence held by Senators in passing Finance related bills with Senators including Sherrod Brown, the Chairman of the Senate Banking Subcommittee on Financial Institutions and Consumer Protection, appearing in the center.

Figure 1: Senate Finance Related Co-sponsorships in the 112th Congress



Source: FiscalNote and S&P Global Market Intelligence Quantamental Research. Data as of 06/09/2021.

1. Introduction

This primer provides a baseline understanding of graph networks using U.S. Senate activity as an example. It is divided into four parts: i) Network Basics, ii) Network Concepts, iii) Network Applications and iv) Creating Networks. Prior papers from S&P Global featuring graph theory analysis have included buy side analyst networks[1], and patent similarity networks[2].

2. Network Basics

A **network** (or **graph**) is a collection of relationships. Network graphs are constructed of two types of objects, known as **nodes** (or **vertices**) and **edges**. Networks are collections of these nodes and edges, and a network must have a minimum of two nodes and one edge.

2.1 Nodes

A node is a component that makes up a network. A node represents an entity or object. Nodes are often described using nouns, for example a company, a student, or a Senator. Nodes can have attributes depending on the use case that can be used to differentiate nodes from each other. For example, a company could have a ticker, a student could have an identification number, or a Senator could have a state.

2.2 Edges

An edge is the second core entity in a network. An edge represents a **relationship** between two nodes. A relationship represents a connection or interaction between two entities and can often be described using verbs. A company can *purchase* products from another company, a student can *study* with another student, and a Senator can *sponsor a bill* with another Senator. To facilitate examples this format will be used when describing edges: "node *relationship* node".

Direction

Edges are further defined by the node at which the edge starts and ends. These are called **source** and **target** nodes. An edge must have a source and target node and cannot have more than one of each. Source and target nodes can also be called **parent** (source) and **child** (target) nodes. The description of an edge can now be written as "source *relationship* target"

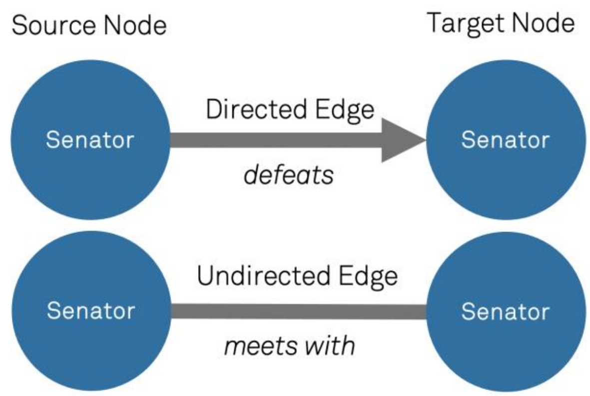
An edge where the order of the source and target nodes is meaningful is known as a **directed** edge. For example, the purchaser and the subsidiary in an acquisition. The edge could be described as the "purchaser *buying from* seller", and it would be incorrect to describe the edge as "seller *buying from* purchaser".

An edge where the order does not matter is an **undirected** edge. The edge "Senator Smith *sponsors a bill with* Senator Jones" is as valid as the edge "Senator Jones *sponsors a bill with*

[1] Oyeniyi et al., (2020)
[2] Rana (2021)

Senator Smith" Another way to describe an undirected edge is **bidirectional**, in that each node can be both the source and the target of the edge.

Figure 2: Directed and Undirected Edges



Source: S&P Global Market Intelligence Quantamental Research.

Weight

Edges can have attributes just like nodes. There is one attribute that is assigned a special significance, **weight**. The weight assigns an importance to the edge in order to compare it with other edges. The weight of an edge is often used when visualizing a network, as in Figure 1, where the weight of the edges was considered when calculating the position of the nodes.

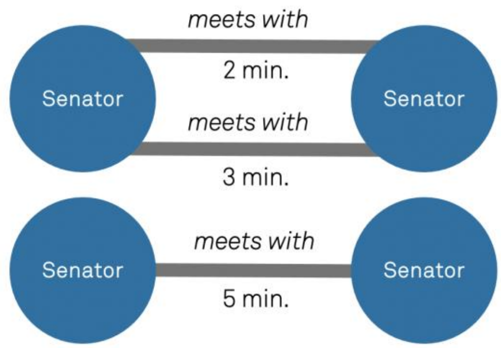
The edge "purchaser *buying from* seller" could have the weight 'price' to allow comparison with edges that represent other transactions. The edge "Senator Jones *sponsors bills with* Senator Smith" could be assigned 'number of bills' as a weight to identify the magnitude of the relationship. The description of an edge can be updated to "source *relationship* target <u>weight</u>"

Duplicate Edges

A graph in which there can be more than one edge with the same source node and target node is called a **multigraph.** For example, if two Senators cosponsored bills in different sessions, there might be two edges, for example "Senator Jones *sponsors bills with* Senator Smith <u>2</u>" and "Senator Jones *sponsors bills with* Senator Smith <u>3</u>". This would indicate that the relationship happened twice, with a different magnitude each time.

Just as tabular data can be aggregated, network edges can also be aggregated. The two previous edges describing Senatorial interactions could be aggregated on the weight, resulting in a single edge, "Senator Jones *sponsors bills with* Senator Smith <u>5</u>." This will result in one edge between each node pair and is often easier to analyze and visualize.

Figure 3: Aggregating Edges



Source: S&P Global Market Intelligence Quantamental Research.

2.3 Network Concepts

Networks, nodes, and edges have attributes created by their interactions.

Adjacency

When an edge describes a relationship between two nodes, the two nodes are said to be **adjacent**, shown in yellow in Figure 4. The **neighborhood** (i.e., neighbors of a node) is the set of all nodes that are adjacent to that node.

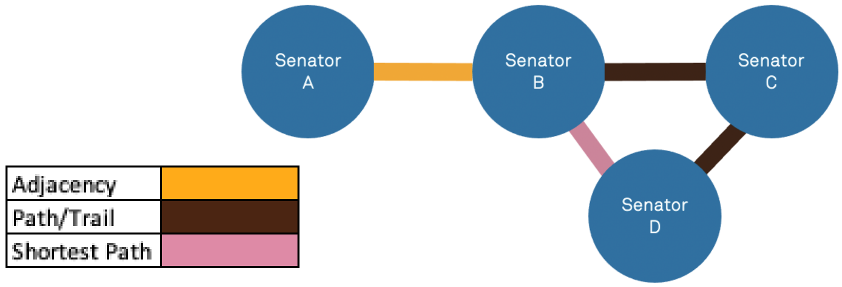
Path

A **path** is the route traveled when moving through the graph from node to node. At each node, there are only as many possible paths as there are adjacent nodes. The **length** of a path is the number of edges traversed when traveling from node x to node y. A path with length 1 is an edge. If the graph is directed, a path can only follow an edge in the direction of source to target. Figure 4 shows a path between Senator B and Senator D in brown.

The **shortest path** is a special path that has the minimum length possible from one node to another. This is shown in pink in Figure 4. The **level** of a node describes the length of the shortest path to a designated **root** node or nodes. The **eccentricity** of a node is the longest possible path to any other node without repeating.

Paths can be used to calculate **betweenness centrality**, which is an attribute of a node that indicates how many shortest paths travel through that node. Betweenness centrality is calculated by finding the shortest paths from all nodes to all other nodes, and then counting the most visited nodes. Betweenness centrality generally indicates which nodes have the most influence or are the most critical to the network, but interpretations can vary based on the use case. For example, a Senator who has high betweenness centrality may have more working relationships, and thus may be able to broker more compromises.

Figure 4: Adjacency and Paths



Source: S&P Global Market Intelligence Quantamental Research.

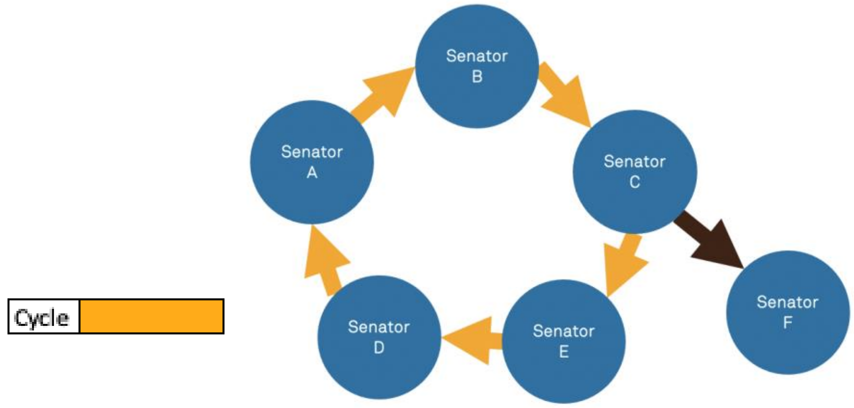
Connectivity

If one node can be reached from another by following a path they are **connected**. If a path can be created that starts and ends on the same node, but does not traverse any other node twice, it is a **cycle**. A cycle in a directed network is shown in Figure 5 in yellow. Cycles can help find loops and other features of the graph that may be of interest to an analyst.

Cycles can be used to divide graphs into **components**, or **subgraphs**. A graph can be constructed from one or more subgraphs. A graph or subgraph in which every node can be reached from every other node through a valid path is **strongly connected.** Directed graphs are rarely strongly connected, and a strongly connected directed network must have at least one cycle. Understanding the structure of a network allows for the appropriate use of network statistics and algorithms as some techniques are better suited for one type or the other. In other use cases strongly connected components can be used to reduce the number of nodes in a network.

A graph or subgraph where every node can be reached from every other node when you can travel forwards or backwards through directed edges is **weakly connected**. A graph that has two or more unconnected subgraphs cannot be weakly or strongly connected. Figure 5 shows a directed network with one cycle that is weakly connected because of the edge in brown. There is no path from Senator F to other nodes in the network, the network could be made strongly connected by adding a node from Senator F to Senator E. If the network depicted here represented email forwards, a message sent to Senators A through E would make it to every other Senator, while a message to Senator F would not.

Figure 5: Cycles and Connectivity



Source: S&P Global Market Intelligence Quantamental Research.

Degree

All nodes have **degree** or the number of edges that arrive or depart from that node. This can be further divided into **in-degree** and **out-degree**, which respectively refer to the number of edges of which the node is a target (in-degree) or source (out-degree).

Intersection

The **intersection** of graph x and graph y is the set of all common nodes between the two networks. This can be used to compare two networks. For example, the intersection of two networks representing the 111th Congress and 112th Congress would show which members were reelected (or not up for reelection).

3. Network Applications

Network graphs can be used to model hidden social relationships[3], employee knowledge within a firm[4], or information dissemination amongst financial analysts.[5] All of these applications (see Appendix B) share a common construction in creating nodes and edges from a series of entities and relationships.

- Business networks are a simple example, where companies serve as nodes with their relationships making up the edges. The economic relationship among firms is often captured in their reports[6] and publications have shown that information about firms disseminates through these networks at a measurable pace.[7]
- Network graphs can map supply chains and production networks. This allows rich analysis of a product's supply chain and components but often requires manual data collection and construction.[8] Supply chain analysis has shifted towards more granular

[3] Huberman et al. (2008)
[4] Parise et al. (2006)
[5] Oyeniyi et al. (2020)
[6] For example, Apple
[7] Cohen & Frazzini (2016)
[8] Nuss et al. (2016)

data that provides a lens into the activity of companies. This lends itself well to a network format, where relationships can stand out to analysts.

- Networks model human relationships well. Social networks are used to uncover important people, influencers, or track the dissemination of knowledge through an organization. For example, a network could be used to identify the most impactful meetings to schedule around a specific topic. A manager might use this to pitch their idea to the most impactful people in a firm, or a lobbyist might identify the lawmakers who interact the most with a specific area of focus.

4. Creating a Network

FiscalNote provides a database of congressional bills and their sponsors. This is provided in a tabular format with each record indicating a bill and the id of one sponsor. It is possible to sort the list by a count of sponsors to answer a question such as "Which senator sponsored the most passed bills in the 112th Congress related to the financial sector?" Appendix C shows the SQL query required to generate this response, with the answer being Senator Lieberman.

A logical follow up question is "Who cosponsored the most bills with Senator Lieberman?" This question, while also simple, is different from the previous query. The first query asked for a summary statistic about an **entity**, while the second asked for a summary statistic about a **relationship.** Tabular data is good at providing rollups on entity but fails to provide an intuitive way of analyzing relationships. Appendix D shows how this can be done using the FiscalNote tables[9]. This additional overhead may be problematic in a larger, more complex database. Analysts who frequently want to query on relationships should consider transforming their data and utilizing a **graph database** to store the transformations. A graph database stores relationship data natively and provides efficiencies in extracting those relationships.

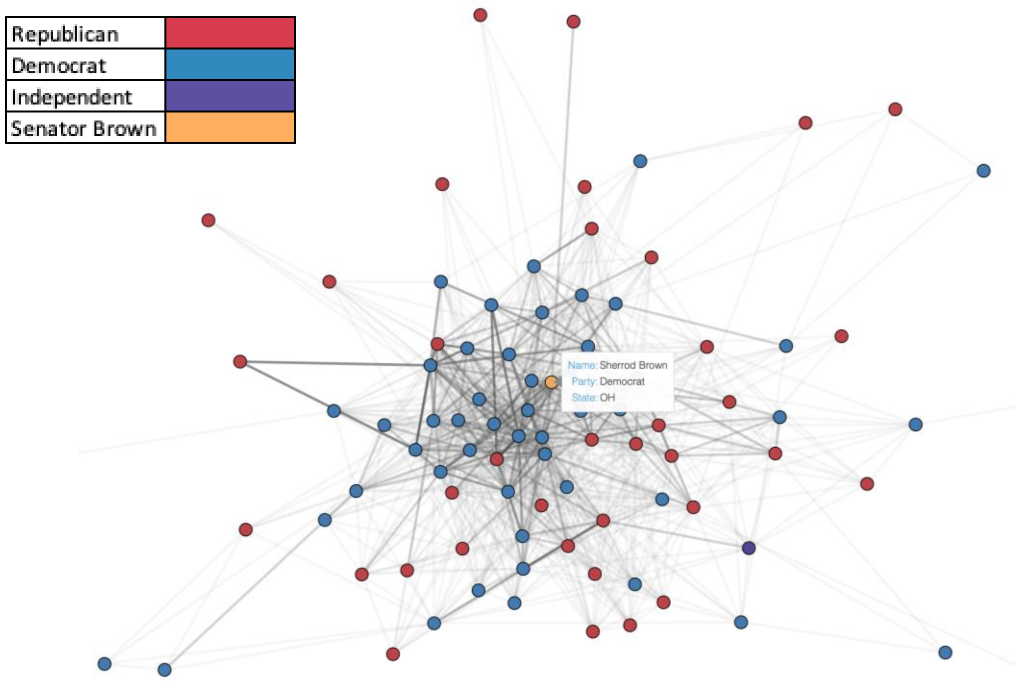
The query in Appendix D creates an **edge list** out of the sponsorship data. This is used to create a simple graph database, illustrating how network concepts come together in practice. The full process is illustrated in Appendix E. The resulting graph contains the relationship data natively and quickly tells us that Senator Lieberman cosponsored the most passed finance related bills in the 112th Congress with Senator Collins, a total of 6.

Another common question is which node is the most important to the network or how to identify nodes of significance. One way to quantify this is using **betweenness centrality**, and the network is structured to quickly provide the answer as illustrated in Appendix E. The node with the highest betweenness centrality is Senator Brown, as highlighted in Figure 6. Senator Brown was the Chairman of the Senate Banking Subcommittee on Financial Institutions and Consumer Protection, it is intuitive that this position would result in high centrality in a graph of finance industry related sponsorships.

[9] Note added complexity and that the table of sponsors needs to be joined to itself, a form of recursion

Data in a network format is simple to visualize. Appendix F illustrates the Python code to transform the edge list generated in Appendix D into a map of nodes and edges. Figure 6 shows how a visualization using a **force directed layout**, which calculates node position with an algorithm that allows nodes to attract along edges and repel other nodes, shows Senator Brown in a central location, visually mirroring the centrality calculation. In the context of finance and business, this central node might represent a key supplier, a company that has considerable analyst coverage, or a manager that coordinates a business-critical team.

Figure 6: Highlighting Senator Brown as a Central Node



Source: FiscalNote and S&P Global Market Intelligence Quantamental Research.

5. Data

Data for the example was drawn from [FiscalNote], a collection of congressional activity available on the S&P Global Marketplace. FiscalNote aggregates laws and regulations from Congress and Federal Agencies in real time, providing information that can be used to assess risk within industries and portfolios. The data includes the full text of legislation from the 111th Congress to present and probability score, industry, topics, sponsors, and more for every bill. Federal rulemaking is also covered with access to Federal Regulations from Agencies and Executive Orders.

6. Conclusion

Alternative and highly granular data incentivize analysts to explore nontraditional methods and tools such as graph networks. This can provide new insights into many domains like supply chain management, equity analysis, and risk preparedness. Networks can be used to quickly identify key objects like the most influential Senator, or to identity important nodes like a key supplier. This allows network analysis to arrive at answers to relationship driven queries directly and to work with relationship data in a more intuitive and human friendly format. Adding network analysis to the discussion of a firm offers insights into the complexity and structure of a firm's business network that cannot necessarily be captured in traditional tabular data and allows humans to quickly identify and qualify outliers and oddities that may be present.

References

Aric A. Hagberg, Daniel A. Schult and Pieter J. Swart, "Exploring network structure, dynamics, and function using NetworkX", in Proceedings of the 7th Python in Science Conference (SciPy2008), Gäel Varoquaux, Travis Vaught, and Jarrod Millman (Eds), (Pasadena, CA USA), pp. 11–15, Aug 2008

Bokeh Development Team (2018). Bokeh: Python library for interactive visualization. http://www.bokeh.pydata.org.

Cohen, L., & Frazzini, A. (2016). Economic Links and Predictable Returns. Journal of Finance, 46.

Huberman, B. A., Romero, D. M., & Wu, F. (2008). Social networks that matter: Twitter under the microscope. 9.

Nuss, P., Graedel, T. E., Alonso, E., & Carroll, A. (2016). Mapping supply chain risk by network analysis of product platforms. Sustainable Materials and Technologies, 10, 14–22. https://doi.org/10.1016/j.susmat.2016.10.002

Oyeniyi, T., Yang, Z., & Tortoriello, R. (2020). The Analyst Matrix—Profiting from Sell-Side Analysts' Coverage Networks. S&P Global Market Intelligence.

Parise, S., Cross, R., & Davenport, T. (2006). Strategies for preventing a knowledge-loss crisis. Sloan Management Review, 47.

Rana, A. (2021). Technology Momentum: Building Peer Networks from Company Patents. S&P Global Market Intelligence

Appendix A: Glossary

Adjacency – a node that shares an edge with another node.
Adjacency matrix – a matrix of length and width n where n is the number of nodes in the network, where each position in the matrix is labeled with a 1 or 0 if an edge exists.
Betweenness Centrality – a score indicating the number of shortest paths that travel through a node.
Bidirected – see undirected.
Child – see target.
Component – see subgraph
Cycle – a path that starts and ends on the same node with no other repeating nodes.
Directed – an edge in which the order of the source and target nodes matters or encodes information.
Eccentricity – the longest possible path from a node to any other node.
Edge – a relationship between a target node and a source node.
Edge List – a method of representing a network as a series of records.
Entity – an object often represented by a noun.
Graph – see network.
Graph Database – a database that stores relationships as source target pairs instead of records related by foreign keys.
In-Degree – the number of edges that list the node as the target.
Intersection – the subset of nodes that appear in both graph A and graph B.
Leaf – a node with degree of 1.
Level – the number of edges a path traverses from a designated root node to a given node.
Multigraph – a network that allows duplicate edges between the same source and target nodes, which may have different weights.
Node – also vertex; is a point that represents an object or entity in a network.
Network – also graph, network graph, graph network, a collection of nodes and edges.
Neighborhood – all adjacent nodes to a given node.
Out-Degree – the number of edges that list the node as the source.
Path – traveling through edges from a starting node to reach a destination node.
Parent – see source.
Relationship – see edge.
Root – an arbitrary start of a network. Often a node with in-degree or out degree of zero.
Strongly Connected – All nodes in the network can be reached from all other nodes in the network, respecting the direction of the edges.
Shortest Path – the shortest possible path from one node to another.
Subgraph – a subset of a network's nodes and edges.
Source – the starting node of the edge**.**
Tree – see network, usually a weakly connected directed network.
Target – the ending node of the edge.
Trail – see path.

Undirected – an edge in which the order of source and target nodes does not matter, or the relationship is mutual.
Vertex – see node.
Weakly Connected – all nodes in the network can be reached from every other node in the network, traveling forward or backward through directed edges.
Weight – a usually numeric value that describes the relationship; can be used to aggregate edges.

Appendix B: Example Network Use Cases

Network	Node(s)	Edge(s)
Senate Co-sponsorships	Senators	Cosponsored bills
Corporate Stock Holdings	Companies	Number of similar holdings
Analyst Stock Coverage	Companies	Number of shared analysts
Board Memberships	Companies	Number of shared board members
Patent Overlap	Companies	Number of similar patents
Corporate Knowledge	Employees	Interactions, emails
Supply Chain	Suppliers	Shipments, transactions
Social	People	Friendship
Professional Meetings	People	Shared attendance

Appendix C: Tabular Lookup

```
WITH passed_finance AS (
SELECT b.billkey, b.congresssession
FROM fnbill b
JOIN fnbillversion v ON v.gpoversioncode = b.gpoversioncode
JOIN fnbillindustry bi ON bi.congresssession = b.congresssession
  AND bi.billkey = b.billkey
  AND left(bi.gicscode, 2) = 40 --Financials
  AND bi.industryscore >= .5
WHERE cqversiontype LIKE '%passed%')
SELECT COUNT(DISTINCT bs.billkey) sponsors, bs.sponsorid, cs.sponsorname
FROM fnbillsponsors bs
JOIN fncongresssponsor cs on cs.sponsorid = bs.sponsorid AND cs.congresssession =
bs.congresssession
JOIN passed_finance ps on ps.billkey = bs.billkey AND ps.congresssession =
bs.congresssession
WHERE bs.congresssession = 112
  AND left(bs.billkey, 1) = 'S' --Senate
GROUP BY bs.sponsorid, cs.sponsorname
ORDER BY sponsors DESC
```

Appendix D: Create an Edge List

```
WITH passed_finance AS (
SELECT b.billkey, b.congresssession
FROM fnbill b
JOIN fnbillversion v ON v.gpoversioncode = b.gpoversioncode
JOIN fnbillindustry bi ON bi.congresssession = b.congresssession
  AND bi.billkey = b.billkey
  AND left(bi.gicscode, 2) = 40 --Financials
  AND bi.industryscore >= .5
WHERE cqversiontype LIKE '%passed%'
GROUP BY b.billkey, b.congresssession
), sponsors AS (
SELECT c.sponsorid, b.BILLKEY, b.CONGRESSSESSION
FROM fnbill b
JOIN fnbillsponsors s ON s.congresssession = b.congresssession
  AND s.billkey = b.billkey
  AND s.gpoversioncode = b.gpoversioncode
JOIN fncongresssponsor c ON c.congresssession = s.congresssession
  AND s.sponsorid = c.sponsorid
JOIN passed_finance pf ON pf.billkey = b.billkey
  AND pf.congresssession = b.congresssession
WHERE b.congresssession = 112
  AND LEFT(b.billkey, 1) = 'S' --Senate
  AND c.statecode NOT IN ('AS', 'FM', 'GU', 'MH', 'MP', 'PR', 'PW', 'VI', 'DC')
GROUP BY c.sponsorid, b.BILLKEY, b.CONGRESSSESSION
), agg AS (
SELECT s1.sponsorid AS sponsor1, s2.sponsorid AS sponsor2, s1.billkey
FROM sponsors s1
JOIN sponsors s2 ON s1.billkey = s2.billkey
  AND s1.sponsorid != s2.sponsorid
  AND s1.sponsorid < s2.sponsorid
)
SELECT sponsor1, cs1.sponsorname, cs1.politicalparty, cs1.statecode,
  sponsor2, cs2.sponsorname, cs2.politicalparty, cs2.statecode,
  COUNT(DISTINCT billkey) AS cosponsors
FROM agg
JOIN fncongresssponsor cs1 ON cs1.sponsorid = sponsor1
  AND cs1.congresssession = 112
JOIN fncongresssponsor cs2 ON cs2.sponsorid = sponsor2
  AND cs2.congresssession = 112
GROUP BY sponsor1, cs1.sponsorname, cs1.politicalparty, cs1.statecode,
  sponsor2, cs2.sponsorname, cs2.politicalparty, cs2.statecode
ORDER BY cosponsors DESC
```

Appendix E: Modeling a Network[10]

To model a network in Python, first start by defining a node, which requires an id. In this case, the node wants to have the attributes name, party, and state. These are defined in a separate class method so that if they are not available the node can still be initiated. The graph object itself is a simple container containing a node object and an edge object. The node object is structured as a dictionary to allow lookups, and an extension to this model could be do redefine the edge object as a dictionary of source, target keys. This would allow lookups on edges. The toy model is then created by calling the Graph class.

```
# define node object
class Node:
   def __init__(self, id):
      self.id = id
      self.has_attrs = False

   def __str__(self):
      return self.id

   def __repr__(self):
      return self.id

   def bulk_set_attrs(self, name, party, state):
      self.name = name
      self.party = party
      self.state = state
      self.has_attrs = True
      return self


# define graph object to serve as a toy database
class Graph:
   def __init__(self):
      self.nodes = {}
      self.edges = []

# start the graph
G = Graph()
```

The edge list generated in Appendix D was saved to a csv file called “fncosponsors.csv”. This then accessed and a list of raw edges created. The script then iterates through the edge list and does threw operations: 1) if the source node is not in the node dictionary, it adds the node to the node dictionary, along with the node attributes assigned in the edge list. 2) if the target node is not in the node dictionary, the target node and attributes are added. 3) the edge is added to the edge list using a reference to the source and target nodes in the node dictionary.

```
# open the edge list and create a list of edges
with open('fncosponsors.csv') as edge_list:
```

[10] This example is also available at: https://github.com/S-P-Quantamental/Branching-Out-Graph-Theory-Fundamentals

```
    edges = [edge.replace('\n', '').split(',') for edge in edge_list]

# iterate through the edge list and add edges to our dictionary. add nodes to our node list if
they do not already exist, with attributes
for edge in edges[1:]:
    if edge[0] not in G.nodes:
        G.nodes[edge[0]] = Node(edge[0]).bulk_set_attrs(name=edge[1], party=edge[2],
state=edge[3])
    if edge[4] not in G.nodes:
        G.nodes[edge[4]] = Node(edge[4]).bulk_set_attrs(name=edge[5], party=edge[6],
state=edge[7])
    source = G.nodes[edge[0]]
    target = G.nodes[edge[4]]
    G.edges.append((source, target, edge[8]))
```

The toy model can now be queried. The first query asks for the largest edge, indicating the largest number of co-sponsorships.

```
# ask the graph object which edge is the largest, and print a response
result = sorted(G.edges, key=lambda x: x[2], reverse=True)[0]
print(f'{result[0].name} ({result[0].party[:1]}-{result[0].state}) and {result[1].name}
({result[1].party[:1]}-{result[1].state}) shared {result[2]} cosponsors')
print('-' * 70)
```

Output:
Joseph I. Lieberman (D-CT) and Susan Collins (R-ME) shared 6 cosponsors
----

The next query asks for a list of all Senators that cosponsored a bill with Senator Lieberman.

```
# ask the graph object for a list of who else cosponsored a bill with Senator Lieberman
result = [y[1] for y in filter(lambda x: x[0].name == 'Joseph I. Lieberman', G.edges)] + [y[0] for
y in filter(lambda x: x[1].id == 'S210', G.edges)]
print('All Senators that cosponsored a bill with Senator Lieberman')
print([x.name for x in result])
print('-' * 70)
```

Output:
All Senators that cosponsored a bill with Senator Lieberman
['Susan Collins', 'Thomas R. Carper', 'Bob Casey', 'Daniel K. Akaka', 'Bill Nelson', 'Jon Tester', 'Mark Pryor', 'Chris Coons', 'Richard J. Durbin', 'Jon Kyl', 'Scott P. Brown', 'Benjamin L. Cardin', 'Jeanne Shaheen', 'Mary L. Landrieu', 'Richard M. Burr', 'Sherrod Brown', 'Amy Klobuchar', 'Mike Lee', 'John Barrasso', 'James M. Inhofe', 'Johnny Isakson', 'Bob Menendez', 'Jeff Sessions', 'Kay Bailey Hutchison', 'Debbie Stabenow', 'Dianne Feinstein', 'Al Franken', 'Marco Rubio', 'Lindsey Graham', 'Pat Roberts', 'Michael B. Enzi', 'John Cornyn', 'Michael Bennet', 'Claire McCaskill', 'David Vitter', 'Mark Begich', 'Kay Hagan', 'Lamar Alexander', 'Charles E. Grassley', 'Carl Levin', 'Tom Harkin', 'Patrick J. Leahy', 'John McCain', 'Richard G. Lugar', 'Orrin G. Hatch', 'Barbara A. Mikulski']

To compute a metrics, the graph is converted to a networkx object to save time reimplementing network algorithms. This is done by creating modified lists out of the node and edge objects and passing them to networkx. Note that networkx can use custom defined objects, like the Node class, as nodes.

```
# compute node statistics using networkx
import networkx as nx

# convert our toy model into a networkx graph
Gnx = nx.Graph()
Gnx.add_edges_from([(s, t, {'cosponsors': int(v)}) for s, t, v in G.edges])
Gnx.add_nodes_from([node for node in G.nodes.values()])
G = Gnx
```

Betweenness centrality can then be computed using the convenience method in networkx. Two examples show how to find the node with the highest betweenness centrality and how to calculate the average centrality by party.

```
# which node is the most central (betweenness)?
cent = nx.betweenness_centrality(G, normalized=True)
cent = ([(i, v) for i, v in cent.items() if not isinstance(i, tuple)])
cent.sort(reverse=True, key=lambda x: x[1])
print(f'Node {cent[0][0].id}, {cent[0][0].name} ({cent[0][0].party[:1]}-{cent[0][0].state}) has the highest betweenness centrality of {cent[0][1]}')
print('-' * 70)

# which party had the highest average betweenness centrality?
party = ([(i.party, v) for i, v in cent])
for p in ('Democrat', 'Republican', 'Independent'):
    tmp = filter(lambda x: x[0] == p, party)
    tmp = [v for i, v in tmp]
    print(f'The average centrality of {p} is {round(sum(tmp) / len(tmp), 4)}')
print('-' * 70)
```

Output:
```
Node S307, Sherrod Brown (D-OH) has the highest betweenness centrality of 0.043802194470251514
----------------------------------------------------------------------
The average centrality of Democrat is 0.0143
The average centrality of Republican is 0.0046
The average centrality of Independent is 0.0052
----------------------------------------------------------------------
```

Appendix F: Visualizing a Network

This example shows how to visualize a network stored as a csv edgelist. The same edge list from Appendix D is used, but it Is ingested directly to networkx instead of passing through a toy model. Most of the code below related to the visual

```
# Load edge list from csv and
# visualize the network using Bokeh
import pandas as pd
import networkx as nx
from bokeh.plotting import figure, from_networkx
from bokeh.models import (Circle, MultiLine)
from bokeh.io import show
from bokeh.palettes import Spectral11, Greys3
from bokeh.models import HoverTool, TapTool, EdgesAndLinkedNodes,
NodesAndLinkedEdges, NodesOnly

# inititalize plot
plot = figure(x_range=(-4, 4), y_range=(-4, 4), toolbar_location="right",
sizing_mode="scale_both")

# remove grid lines and axis
plot.xgrid.grid_line_color = None
plot.ygrid.grid_line_color = None
plot.axis.visible = False
```

Node an edge attribute is calculated before creating the graph, this is author preference

```
# initialize networkx graph from csv edgelist
df = pd.read_csv('fncosponsors.csv')
df.columns = ['source', 'source_name', 'source_party', 'source_state', 'target', 'target_name',
'target_party', 'target_state', 'cosponsors']
df.fillna('', inplace=True)
df['edge_fill'] = df['cosponsors'].apply(lambda w: ((w - df['cosponsors'].min()) /
(df['cosponsors'].max() - df['cosponsors'].min()) + .05))
G = nx.from_pandas_edgelist(df, 'source', 'target', ['cosponsors', 'edge_fill'], nx.Graph)

# tooltips
node_hover = HoverTool(line_policy='interp', tooltips=[("Name", "@name"), ('Party',
'@party'), ('State', '@state')])
plot.add_tools(node_hover, TapTool())
```

Dictionaries can be created that map attributes to colors. Load these into the network as additional attributes to designate styling by node.

```
# create custom color mappings
party_map = {'Republican': Spectral11[9], 'Democrat': Spectral11[1], 'Independent':
Spectral11[0]}
```

Node attributes are set using a loop over the edge list. The source node and target node data are validated and appended. This implementation assumes that every node attribute instance is identical.

```
# set node attributes
node_atrs = {}
for i, row in df.iterrows():
    if row['source_party'] != '':
        node_atrs[row['source']] = {'name': row['source_name'], 'party': row['source_party'],
'state': row['source_state'], 'node_color': party_map[row['source_party']]}
    if row['target_party'] != '':
        node_atrs[row['target']] = {'name': row['target_name'], 'party': row['target_party'], 'state':
row['target_state'], 'node_color': party_map[row['target_party']]}
nx.set_node_attributes(G, node_atrs)

# create layout - customization and interactive functions
layout = from_networkx(G, nx.spring_layout, scale=1, center=(0, 0), weight='cosponsors')
layout.node_renderer.glyph = Circle(size=15, fill_color='node_color')
layout.node_renderer.selection_glyph = Circle(size=15, fill_color='node_color')
layout.node_renderer.hover_glyph = Circle(size=15, fill_color=Spectral11[7])
layout.edge_renderer.glyph = MultiLine(line_color=Greys3[0], line_alpha='edge_fill',
line_width=3)
layout.edge_renderer.selection_glyph = MultiLine(line_color=Spectral11[7], line_width=3)
layout.edge_renderer.hover_glyph = MultiLine(line_color=Spectral11[6], line_width=3)
layout.selection_policy = NodesAndLinkedEdges()
layout.inspection_policy = NodesOnly()
plot.renderers.append(layout)

show(plot)
```

The visualization output is also available as an HTML file at: https://github.com/S-P-Quantamental/Branching-Out-Graph-Theory-Fundamentals

Our Recent Research

May 2021: U.S Filings: No News is Good News

Company annual filings are a vital but often under-analyzed source of information for investors. Market moving content is buried within an ever-growing body of text that on average is equivalent to a 240-page novel. The filings contain subtle revisions making a computational linguistic approach imperative. Faced with this voluminous amount of text and the minute number of changes, investors have historically overlooked the newly embedded information and the implications of those additions

March 2021: Hiding in Plain Sight – Risks That Are Overlooked

This report uses three metrics (Minimum Edit Distance, Jaccard Similarity, and Cosine Similarity) to identify companies that made significant changes to the "Risk Factors" section of their filings. These metrics can serve as alpha signals or be used to quickly identify a pool of companies that require further investigation.

January 2021: Leadership Change That Matters: A Value and Momentum Story

December 2020: Warranted Optimism: Sentiment vs. Supply Chain

December 2020: A Dark Winter for REITS: Trouble Brewing

Add another item to the list of 'firsts' for 2020 - markets saw two REITs declare bankruptcy on the same day for the first time. CBL Properties (CBL) and Pennsylvania Real Estate Investment Trust (PEI) filed for bankruptcy protection on Monday, November 2, 2020. An examination of those factors common to CBL and PEI reveals similar operational and fundamental characteristics. Our analysis reveals seven REITs that share these common characteristics: 1) a high percentage of anchor tenants that have declared bankruptcy; 2) a decline in building permit activity, 3) a decline in foot traffic, 4) a high degree of leverage, 5) declining cash flow, and 6) a high proportion of tenants that have filed for bankruptcy.

October 2020: Sweet Spots in the C-Suite: Executive Best Practices for Shareholder Friendly Firms

The Business Roundtable, an association of CEOs of America's leading companies, published a new statement on corporate responsibility in August 2019. The statement identifies five important corporate stakeholders: customers, employees, suppliers, communities and shareholders.1 This report highlights four key types of executive policy that drive value creation for stakeholders: profitability vs. growth decisions, mergers & acquisitions policy, return of cash to shareholders, and insider stock ownership. In it, we demonstrate empirically those practices that increase corporate value over time, thereby rewarding shareholders, employees, and other stakeholders. These practices also form a scorecard by which stakeholders can evaluate whether or not management is undertaking actions likely to increase corporate prosperity.

October 2020: Just the (Build)Fax: Property Intelligence from Building Permit Data

Investors have struggled to price the uncertainties facing commercial Real Estate owners as tenants' businesses continue to be impacted by the economic lockdown following the COVID19 global pandemic. Building permit1 data has provided valuable transparency for investors of Real Estate Investment Trusts (REITs) during this time. Publicly traded REITs that have continued permit-requiring activities have outperformed their peers during and following the economic lockdown. A one standard deviation increase in a derived permit ratio, compared to the industry2 average, is associated with 427 bps of excess return over the same period.

August 2020: <u>The Analyst Matrix: Profiting from Sell-Side Analysts' Coverage Networks</u>
Sell-side analyst coverage data provides a new and rich source of establishing connections between firms, as analysts (given their industry expertise) are likely to cover fundamentally related firms. This report uses sell-side analysts' coverage data to build a connected-firm network (CFN) - a portfolio of companies that are covered by analyst(s) that follow a focal firm. This network has three broad applications: measuring the "strength" of economic relationships between companies; forecasting fundamentals of companies in the network; and as a stock selection signal.

June 2020: <u>The Information Supply Chain Begins Recovering From COVID</u>

The COVID-19 shockwaves emanating through the global supply chain continue to reverberate. The information that decision makers have traditionally relied on have also been disrupted but is slowly showing signs of normalizing. S&P Global Market Intelligence processes 64,000 financial documents each day, placing it in a central position in the information supply chain with a unique view into the specific areas and magnitude of information disruption.

May 2020: <u>Never Waste a Crisis: Following the Smart Money Through Beneficial Ownership Filings</u>

Investors looking for ideas amid the recent market downturn may profit from reviewing beneficial ownership filings: SEC schedules 13D and 13G. These purchases often represent high conviction buys by activists, industry insiders, hedge funds, etc. Our previous investor activism research shows that investors can benefit by following activists' lead: a portfolio of stocks that activists had targeted outperformed the market by over 8% annually.

This report examines recent 13D and 13G filings, and spotlights four purchases of target companies with high historical operating cash flows and below average dividend payments, characteristics of companies typically targeted by activists.

May 2020: <u>Risky Business: Foot Traffic, Vacancy Rates and Credit Risks</u>

The COVID-19 pandemic has led to widespread closures of retail stores, offices and hotels. Foot traffic data can be combined with traditional financial ratios to provide a more holistic view of business health for both credit and equity investors. This report extends our prior analysis of foot-traffic data by setting foot traffic figures in the context of a screen for identifying where risks may be highest.

The analysis in this report can help: i) Creditors identify customers that require additional credit facilities to support growth, or companies where existing credit lines need to be reassessed given bleak prospects; and ii) Equity investors identify companies where revenues may be accelerating or firms that may have difficulty meeting financial obligations.

May 2020: Finding the Healthy Stocks in Health Care During Lockdown

Elective and non-essential medical procedures are on an indefinite hold in many places. Simultaneously, essential medical services are in high demand, and likely to remain in demand for the near future. This dynamic creates winners and losers among Health Care device manufacturers and distributors. Investors can identify potential opportunities in the Health Care Equipment and Services subsector by analyzing 510(k) premarket notifications, which are filings required by the U.S. Food and Drug Administration (FDA) for any company seeking to market a medical device in the United States.

May 2020: No More Walks in the (Office) Park: Tying Foot Traffic Data to REITs

Foot traffic data provides investors and corporate managers with key insights on the level of activity at properties and the demographic profile of visitors to these locations. Corporate managers can use this information to pinpoint properties at greater risk of tenant defaults, while investors can use foot traffic data to identify REITs managing properties where activity remains robust. More importantly, once the nationwide lockdown eases, foot traffic can serve as a leading indicator of a return of economic activity across industries.

May 2020: Do Markets Yearn for the Dog Days of Summer: COVID, Climate and Consternation

Stakeholders are turning to untraditional data sources to quantify the impact of the COVID-19 shutdown. While no single variable can forecast which locations will be most susceptible to the virus, mounting scientific literature suggests that there is a correlation between temperature and viral propagation. If correct, regions in the temperature 'target zone' may need to implement more stringent lockdown policies for a longer period to achieve comparable mitigation.

Investors can combine weather data with property data, to expose one dimension of risk for Real Estate Investment Trusts (REITs) of prolonged closures, as well as areas that may see a resurgence of the virus later this year.

April 2020: Cold Turkey - Navigating Guidance Withdrawal Using Supply Chain Data

A recent surge in corporate earnings guidance withdrawals has left decision-makers missing a wrench in their toolbox. Corporate guidance was already declining, in 2018, when the number of companies in the Russell 3000 providing guidance peaked at 1,721, dropping 6.9% year over year in 2019 to 1,632 companies. Guidance has been further impacted by the Coronavirus pandemic – 173 companies withdrew their previous guidance in the first quarter.

This leaves decision-makers looking for alternative forward-looking information on a company's prospects.

April 2020: Data North Star - Navigating Through Information Darkness

Crisis creates uncertainty. Familiar landmarks lose their value and decision makers are left to navigate on partial information. Following the outbreak of the COVID-19 pandemic, this is the environment in which investors and corporate decision-makers now suddenly find themselves. The S&P Global Quantamental Research team has launched a series of research briefs that will aid decision-makers in navigating this uncertain environment. Utilizing non-traditional datasets across the entire S&P Global Market Intelligence product suite, these briefs will provide market participants with analysis on COVID-19's impact to the financial markets geared to fill the current information gap.

March 2020: Long Road to Recovery: Coronavirus Lessons from Supply Chain and Financial Data

COVID-19 continues to disrupt global supply chains in unprecedented ways. Leveraging maritime shipping data from Panjiva, this report includes a review of trade and financial data to analyze the impact of the SARS-CoV-2 / COVID-19 coronavirus outbreak. Findings include:

- Second-order supply chain effects are also emerging with the apparel industry now seeing a shortage of materials globally due to earlier outages in China.
- Retailers including Costco and Target are gaining from increased sales of health- and personal care products. Yet, supply shortages are rapidly emerging in part due to medical supply export restrictions in several countries.
- There is a notable, but not statistically significant, relationship with firms with higher exposure to Asia having seen a weaker sector neutral stock price performance.

February 2020: Ship to Shore: Mapping the Global Supply Chain with Panjiva Shipping Data in Xpressfeed™

World merchandise trade accounted for an estimated $19.7 trillion in 2018, about 90% of which is by sea. While financial data tells us "how a company has done in the past," shipping data provides a closer-to-real time indicator of "what a company is doing now." Panjiva's shipping data allows investors to track trends, identify anomalies, and assess risks for companies engaged in international trade. This paper illustrates how to find investment insights in Panjiva's US seaborne and Mexican datasets using the US auto parts industry as a case study.

Findings include:

- Shipment trends often lead fundamentals: Rising shipments amid flat or declining fundamentals may signal future financial trend reversal
- Growth in the number of a company's suppliers and in the types of products it imports may signal strengthening demand and/or product line diversification.

- Tracking industry-level product-line trends can help identify companies with significant exposure to rising or declining product lines.

January 2020: Natural Language Processing – Part III: Feature Engineering Applying NLP Using Domain Knowledge to Capture Alpha from Transcripts

Unstructured data is largely underexplored in equity investing due to its higher costs. One particularly valuable unstructured data set is S&P Global Market Intelligence's machine-readable earnings call transcripts.

- Topic Identification – Firms that referenced the most positive descriptors around their financials outperformed historically.
- Transparency – Firms that provided greater call transparency exhibited by executives' behaviors and decisions outperformed historically.
- Weighted Average Sentiment – Quantifying call sentiment using a weighted average construct led to better returns and less volatility historically.
- Additive Forecasting Power – The newly introduced signals demonstrated additive forecasting power above commonly used alpha and risk signals historically.

December 2019: The "Trucost" of Climate Investing: Managing Climate Risks in Equity Portfolios

Does sustainable investing come at a "cost", and is the fear of investors around the performance concessions of "green" portfolios warranted? Our latest research suggests investors' fears are misplaced – carbon-sensitive portfolios have similar returns and significantly better climate characteristics than portfolios constructed without carbon emission considerations. Other findings include:

- Highly profitable firms are likely to be leaders in reducing their carbon emission levels.
- There is no degradation in fundamental characteristics for the carbon-sensitive portfolios compared to the baseline portfolio, even though the difference in constituents can be as high as 20%.
- Carbon-sensitive portfolios were observed as having significant reductions in water use, air pollutants released, and waste generated.

October 2019: #ChangePays: There Were More Male CEOs Named John than Female CEOs

This report examines the performance of firms that have made female appointments to their CEO and CFO positions. Our research finds that firms with female CEOs and/or CFOs:

- Are more profitable and generated excess profits of $1.8 trillion over the study horizon.
- Have produced superior stock price performance, compared to the market average.
- Have a demonstrated culture of Diversity and Inclusion, evinced by more females on the company's board of directors.

June 2019: Looking Beyond Dividend Yield: Finding Value in Cash Distribution Strategies

June 2019: The Dating Game: Decrypting the Signals in Earnings Report Dates

May 2019: Bridges for Sale: Finding Value in Sell-Side Estimates, Recommendations, and Target Prices

February 2019: U.S Stock Selection Model Performance Review

February 2019: International Small Cap Investing: Unlocking Alpha Opportunities in an Underutilized Asset Class

January 2019: Value and Momentum: Everywhere, But Not All the Time

November 2018: Forging Stronger Links: Using Supply Chain Data in the Investing Process

September 2018: Their Sentiment Exactly: Sentiment Signal Diversity Creates Alpha Opportunity

September 2018: Natural Language Processing – Part II: Stock Selection: Alpha Unscripted: The Message within the Message in Earnings Calls

July 2018: A Case of 'Wag the Dog'? - ETFs and Stock-Level Liquidity

June 2018: The (Gross Profitability) Trend is Your Friend

May 2018: Buying the Dip: Did Your Portfolio Holding Go on Sale?

March 2018: In the Money: What Really Motivates Executive Performance?

February 2018: The Art of the (no) Deal: Identifying the Drivers of Canceled M&A Deals

January 2018: U.S Stock Selection Model Performance Review

September 2017: Natural Language Processing - Part I: Primer

July 2017: Natural Language Processing Literature Survey

June 2017: Research Brief: Four Important Things to Know About Banks in a Rising Rate Environment

April 2017: Banking on Alpha: Uncovering Investing Signals Using SNL Bank Data

March 2017: Capital Market Implications of Spinoffs

January 2017: U.S. Stock Selection Model Performance Review 2016

November 2016: Electrify Stock Returns in U.S. Utilities

October 2016: A League of their Own: Batting for Returns in the REIT Industry - Part 2

September 2016: A League of their Own: Batting for Returns in the REIT Industry - Part 1

August 2016: Mergers & Acquisitions: The Good, the Bad and the Ugly (and how to tell them apart)

July 2016: Preparing for a Slide in Oil Prices -- History May Be Your Guide

June 2016: Social Media and Stock Returns: Is There Value in Cyberspace?

April 2016: An IQ Test for the "Smart Money" – Is the Reputation of Institutional Investors Warranted?

March 2016: Stock-Level Liquidity – Alpha or Risk? - Stocks with Rising Liquidity Outperform Globally

February 2016: U.S. Stock Selection Model Performance Review - The most effective investment strategies in 2015

January 2016: What Does Earnings Guidance Tell Us? – Listen When Management Announces Good News

November 2015: Late to File - The Costs of Delayed 10-Q and 10-K Company Filings

October 2015: Global Country Allocation Strategies

September 2015: Research Brief: Building Smart Beta Portfolios

September 2015: Research Brief – Airline Industry Factors

August 2015: Point-In-Time vs. Lagged Fundamentals – This time i(t')s different?

August 2015: Introducing S&P Capital IQ Stock Selection Model for the Japanese Market

July 2015: Research Brief – Liquidity Fragility

May 2015: Investing in a World with Increasing Investor Activism

April 2015: Drilling for Alpha in the Oil and Gas Industry – Insights from Industry Specific Data & Company Financials

February 2015: U.S. Stock Selection Model Performance Review - The most effective investment strategies in 2014

January 2015: Research Brief: Global Pension Plans - Are Fully Funded Plans a Relic of the Past?

January 2015: Profitability: Growth-Like Strategy, Value-Like Returns - Profiting from Companies with Large Economic Moats

October 2014: Lenders Lead, Owners Follow - The Relationship between Credit Indicators and Equity Returns

July 2014: Factor Insight: Reducing the Downside of a Trend Following Strategy

May 2014: Introducing S&P Capital IQ's Fundamental China A-Share Equity Risk Model

April 2014: Riding the Coattails of Activist Investors Yields Short and Long Term Outperformance

March 2014: Insights from Academic Literature: Corporate Character, Trading Insights, & New Data Sources

February 2014: Obtaining an Edge in Emerging Markets

February 2014: U.S Stock Selection Model Performance Review

January 2014: Buying Outperformance: Do share repurchase announcements lead to higher returns?

October 2013: Informative Insider Trading - The Hidden Profits in Corporate Insider Filings

September 2013: Beggar Thy Neighbor – Research Brief: Exploring Pension Plans

August 2013: **Introducing S&P Capital IQ Global Stock Selection Models for Developed Markets: The Foundations of Outperformance**

July 2013: **Inspirational Papers on Innovative Topics: Asset Allocation, Insider Trading & Event Studies**

June 2013: **Supply Chain Interactions Part 2: Companies – Connected Company Returns Examined as Event Signals**

June 2013: **Behind the Asset Growth Anomaly – Over-promising but Under-delivering**

April 2013: **Complicated Firms Made Easy - Using Industry Pure-Plays to Forecast Conglomerate Returns**.

March 2013: **Risk Models That Work When You Need Them - Short Term Risk Model Enhancements**

March 2013: **Follow the Smart Money - Riding the Coattails of Activist Investors**

February 2013: **Stock Selection Model Performance Review: Assessing the Drivers of Performance in 2012**

January 2013: **Research Brief: Exploiting the January Effect Examining Variations in Trend Following Strategies**

December 2012: **Do CEO and CFO Departures Matter? - The Signal Content of CEO and CFO Turnover**

November 2012: **11 Industries, 70 Alpha Signals -The Value of Industry-Specific Metrics**

October 2012: **Introducing S&P Capital IQ's Fundamental Canada Equity Risk Models**

September 2012: **Factor Insight: Earnings Announcement Return – Is A Return Based Surprise Superior to an Earnings Based Surprise?**

August 2012: **Supply Chain Interactions Part 1: Industries Profiting from Lead-Lag Industry Relationships**

July 2012: **Releasing S&P Capital IQ’s Regional and Updated Global & US Equity Risk Models**

June 2012: **Riding Industry Momentum – Enhancing the Residual Reversal Factor**

May 2012: **The Oil & Gas Industry - Drilling for Alpha Using Global Point-in-Time Industry Data**

May 2012: **Case Study: S&P Capital IQ – The Platform for Investment Decisions**

March 2012: **Exploring Alpha from the Securities Lending Market – New Alpha Stemming from Improved Data**

January 2012: **S&P Capital IQ Stock Selection Model Review – Understanding the Drivers of Performance in 2011**

January 2012: **Intelligent Estimates – A Superior Model of Earnings Surprise**

December 2011: **Factor Insight – Residual Reversal**

November 2011: **Research Brief: Return Correlation and Dispersion – All or Nothing**

October 2011: **The Banking Industry**

September 2011: **Methods in Dynamic Weighting**

September 2011: **Research Brief: Return Correlation and Dispersion**

July 2011: **Research Brief - A Topical Digest of Investment Strategy Insights**

June 2011: **A Retail Industry Strategy: Does Industry Specific Data tell a different story?**

May 2011: **Introducing S&P Capital IQ's Global Fundamental Equity Risk Models**

May 2011: **Topical Papers That Caught Our Interest**

April 2011: **Can Dividend Policy Changes Yield Alpha?**

April 2011: **CQA Spring 2011 Conference Notes**

March 2011: **How Much Alpha is in Preliminary Data?**

February 2011: **Industry Insights – Biotechnology: FDA Approval Catalyst Strategy**

January 2011: **US Stock Selection Models Introduction**

January 2011: **Variations on Minimum Variance**

January 2011: **Interesting and Influential Papers We Read in 2010**

November 2010: **Is your Bank Under Stress? Introducing our Dynamic Bank Model**

October 2010: **Getting the Most from Point-in-Time Data**

October 2010: **Another Brick in the Wall: The Historic Failure of Price Momentum**

July 2010: **Introducing S&P Capital IQ's Fundamental US Equity Risk Model**