***NIH / USPTO Cancer Network Miner***

***Booz Allen Hamilton and Omnity***

Booz Allen Hamilton and Omnity have partnered to demonstrate the power of data science in addressing the USPTO’s Cancer Moonshot Challenge. Booz Allen Hamilton is a leading provider of management and technology consulting services to the US government in defense, intelligence, and civil markets, and to major corporations, institutions, and not-for-profit organizations. Omnity enables searchers to efficiently find related documents, even if those documents do not directly cite or link to one another, which accelerates the discovery of otherwise hidden, high-value patterns of interconnection.

The ***NIH / USPTO Cancer Network Miner*** will allow the National Institutes of Health (NIH), other funding institutions, and cancer researchers to see how research efforts are linked into networks of similar and complementary research. Our solution shows how individual research institutions can serve as a lynchpin between related groups of researchers, and how knowledge and discoveries are disseminated through research communities. This will help NIH managers better understand the return on investment from research grants by revealing otherwise hidden conceptual connections shared between patents, research grants, and clinical trials. The USPTO can rapidly extend this approach to a larger and deeper document set and enrich it with additional capabilities.

**1. The Cancer Research Landscape: A Self-Assembling Community Analysis**

The inspiration for our solution was the observation that in the data provided by the USPTO, many patents were supported by one or more NIH grants, and that many grants supported numerous patents. Using these relationships, we built a network graph where each node is a patent, and each edge is a grant linking two patents together. In the graph, each patent appears only once, while a single grant may appear many times as it supports numerous patents.

The graph shows how patents and grants self-assemble into a series of clusters, or research communities, which can represent a range of organizational scale, from a sprawling research network to a single patent and grant pairing. In all, 1,248 distinct clusters were identified.



Figure A single cancer network cluster.

The strength of the relationship between linked patents is described by the thickness of the lines connecting them. The line thickness represents the degree of semantic similarity between patent documents, 1 being identical, and 0 being completely different, based on data provided by Omnity. Figure 1 is a single cancer network cluster showing the relationships between 6 research institutions and 4 NIH Institutes. The thickness of the lines show the semantic similarities between patents. Our analysis showed that the semantic relationship is real, and decays with increased network distance (Figure 2).

This framework allows stakeholders to rapidly detect where the greatest return on investment exists within an otherwise complex landscape of patent families. Rolling a cursor over a patent brings out a pop-up box with information about the cluster (identification number, total funding[[1]](#footnote-1), and dollars per patent produced), and information about the patent (grant recipient organization, patent title, abstract[[2]](#footnote-2) and patent ID).

Figure Semantic Distance vs Network Distance for patents.

**2. A timeline of Semantically Interconnected Patents, Research Grants, and Clinical Trials**

Our solution allows the user to drill down into information on each patent within the network graph. Clicking on a patent node opens a second browser window to reveal a semantic analysis of patents, research grants, and clinical trials most closely related to the patent (Figure 3). This Omnity-developed linguistic capability detects related concepts between documents based on shared, unique words. Following connections through shared concepts, the user can identify how patents, pending patents, grants and clinical trials have evolved over time to further understand the research landscape.



Figure The Omnity exploration tool.

***3. A Framework for Actionable Insights***

Our solution provides the USPTO with valuable, actionable information. Understanding which patents are linked to the greatest amount of research funding and the largest clinical trials, USPTO management can then expedite examination of these patent classes in cancer research to reduce their time to market and accelerate their impact. Additionally, stakeholders within the research community can better understand what cancer technologies have been most impactful.

**4. Accelerating Cancer Diagnostics and Therapeutics**

Accelerating patent prosecution will provide biotech, medical device, pharmaceutical, life sciences and healthcare companies more certainty about their patent applications, reducing speculation about the business value of their investments while increasing confidence in the protection of their intellectual property. More certainty will accelerate decisions, which in turn will lead to earlier clinical trial progress for cancer drugs.

**5. Rapidly Identifying Highly Productive Research Areas at Scale**

The USPTO can further use our solution to demonstrate to key stakeholders at the NIH the value that cancer research provides to the scientific and medical community. Much of the innovation related to cancer is interdisciplinary in nature – our solution illuminates the research domains that are driving innovation to help identify highly productive domains. Our dashboard can connect researchers across research institutes with one another. These connections can drive broader engagement and greater information sharing with grantees, whose individual innovations may be synergistically associated with a cluster or a funding network.

We believe that the connections uncovered in our solution provide the USPTO and its partners with a framework to rapidly and efficiently address several key aspects of public policy formulation that will help accelerate our collective objective of eradicating cancer. Through better allocation of its own resources, facilitating connections within the research community, and demonstrating the positive externalities of cancer research, the USPTO can drive the Cancer Moonshot initiative to a successful conclusion.

**NIH / USPTO Cancer Network Miner**

**Testing Instructions**

*Direct questions to Dr. Thomas Carson at Booz Allen Hamilton by email (carson\_thomas@bah.com) or phone 240-344-1744.*

The NIH / USPTO Cancer Network Miner is a network visualization where every node is a patent, and every edge is an NIH grant that links two patents together by common funding. It is intended to help explore relationships within the cancer research community, and enable deeper insights about the links between various research institutions and NIH funding institutes.

# **1. Open Link and Sign In**

Users can link to our submission through Tableau Online using the following link:

goo.gl/Jps4Kl

Users can use the following credentials to access the link

**Login:** USPTOcancermoonshot@gmail.com

**Password:** Moon2016

We have some suggestions to work with Tableau Online or your web browser:

* ***We recommend users access the link using Internet Explorer on a Windows machine, and Firefox on a Mac.***
* ***The first time you select a patent node to activate Omnity (Step 10), your browser may block a pop-up. You will need to allow pop-ups from that source.***
* ***Hovering over a specific legend entry at the bottom of the Patent Grants Network Map should bring up the full name of the entry (ie. NIH Grant Recipient or NIH Funding Institute). For some reason, Tableau Online does not always bring up the proper name, but rather the name below it.***

# **Note Visualization Overview in the Lower Right Corner**

This text box provides a high-level overview of the visualization, and enough instructions for the user to start exploring the data.

# **Select “Number of Clusters Shown on Map”**

The visualization defaults to 20 clusters visible on opening. Through a drop-down menu, the user can select “All, 5, 20, 50, 100, or 500” clusters to display in the map. The clusters are shown based on the highest number of patents in the cluster. Reducing the number of clusters makes it easier to compare the structure of the clusters. Note that the clusters take on different forms. Some are tightly packed, having many patents, but only a few different grants and NIH funding institutes. Others are sprawling, linking many research institutions and funding institutes.

# **Navigation within the “Patent Grant Networks Map”**

Moving the cursor into the map activates the navigation tools on the upper left side of the map. These allow the user to zoom, select and scroll through the map. If you get lost, the home button on the tool will restore the view to show all visible entities on the map.

# **Activate Information Boxes for Each Node and Edge**

Hovering your cursor over a specific circle (node) in the network map will bring up an information box showing information about the cluster that the patent belongs to, as well as information on the specific patent. Hovering the cursor over a line (edge) in the network map will bring up an information box showing information about the specific grant represented by that edge.

# **Select Specific Network Cluster from Network Clusters Table**

The ***Network Clusters*** table in the upper-right of the visualization lists the Network Cluster number, the Cluster Total Funding, and the number of patents or publications in the cluster. The table is sorted by the network cluster funding. By clicking on one of the network cluster numbers, the visualization focuses on the single cluster. The cell with the cluster number darkens. The ***Patent Grant Networks Map*** zooms to the single cluster, and the ***Patents by NIH Institute and Activities*** and ***Network Cluster Technology Summary*** graphics are populated. These graphics are only populated when a single cluster is selected, because otherwise they are too densely packed to be legible.

We recommend that users conduct their analysis by selecting a single cluster. The user will see that the circles (patents) are colored according to the research institution granted the patent, and the lines colored by the NIH Funding Institute. The size of the lines connecting patents is indicative of the degree of similarity of the texts of the patents that they are connecting. A thicker line indicates a high degree of similarity between the patents, while a thinner line indicates a lower degree of similarity. The semantic similarity is a number between 0 and 1, where 1 is a perfect match. Hovering your cursor above one of the lines brings up an information box about the grants, as well as the **Patent-Patent Semantic Match** number. The semantic match numbers were provided by ***Omnity***.

By exploring within a single cluster, the relationships between grants and funding institutions start to become clear. Often, multiple funding institutes will contribute to a single grant, and single grants may fund many patents. In some areas of the cluster, grant lines may be very thick, indicating that the patents in that area are very similar, while in other areas, it can be seen that the patents related by a grant are not substantially similar.

*The network cluster can be* ***unselected*** *by clicking on the cluster number a second time.*

# **Examine the Patents by NIH Institute and Activities Graphic and the Network Cluster Technology Summary Graphic**

The ***Patents by NIH Institute and Activities*** graphic and the ***Network Cluster Technology Summary*** graphic are populated with information about the selected cluster whenever a cluster is selected. The ***Patents by NIH Institute and Activities*** graphic allows the user to see which NIH institutes and activity codes are contributing grants to that cluster, and how many patents are associated with that institute in the cluster. The ***Network Cluster Technology Summary*** graphic shows which technologies (of the 9 listed in the USPTO Cancer data file based on CPC codes) are used in the cluster. Scrolling over the bars in the charts brings up additional information about the data.

# **Highlight Specific NIH Grant Recipient and NIH Funding Institutes in the Patent Grant Network Map**

The color legend below the ***Patent Grant Network Map*** identifies that the nodes (circles) are colored according to the ***NIH Recipient Organization*** (research institute). The edges (lines) are colored by the ***NIH Funding Institute***. Clicking on one of these entities in the legend will highlight that entity in the ***Patent Grant Network Map,*** making it easier to locate them.

***Clicking within whitespace on the map will deselect that entity.***

# **Filter Network Map to Only Clusters that Contain Patents with Successful FDA Applications**

To start this, make sure you have ***unselected*** the network clusters, so that all are visible in the map. Next, set the ***Number of Clusters Shown on Map*** to ***All***. The number of patents associated with federal grants and successful FDA Applicants is small. Next, in the ***Only FDA Applicants in Map*** control, select ***Yes***. This will filter the ***Patent Grant Network Map*** down to the seven clusters in this category. Selecting ***No*** in the ***Only FDA Applicants in Map*** will return the map to its normal state.

This allows the user to inspect those rare clusters that contain patents that had both NIH Funding and FDA drug trials (according to the USPTO Cancer 12A data set). In all, there are only about 25 patents in the data set that are in both categories.

# **Reveal Conceptual Connections using Omnity’s Semantic Analysis**

Clicking on a node (circle) in the ***Cancer Network Miner*** will open up a second web browser tab, where the user can explore Omnity’s time series visualization in order to move beyond the patents presented in the networked visualization and dig deeper into the patents of interest.

Omnity’s time series visualization displays information about the specific patent selected as well as other patents, grants, and clinical trials that are considered semantically similar, based on rare shared words in the full text of each document.  Documents are plotted by their organization or author on a timeline.  This interactive display allows users to explore the semantic connectivity between documents (*those sharing the same concept or idea*) over time.

Organizations are shown on the left-hand side of the screen.  Documents are shown in the center visualization, and information about the currently selected document is shown on the right-hand side of the screen in the *info panel*.

The documents shown in the visualization are authored by the organizations selected on the left-hand side of the screen.  Select from the list of organizations to show documents authored by that selection in the visualization. Click on an single organization to see its documents and all related documents.  Click a document node to see information about that document in the *info panel*.

In the visualization, the Y (vertical) axis shows the organization corresponding to the document.  The X (horizontal) axis shows the year applicable to the document (e.g. patent application filed, patent granted, clinical trial [completion date or start date, or first received], federal grant posted).  The color legend above the visualization identifies that the nodes (circles) are colored according to the document source type - federal grants in green, clinical trials in blue, and patents in yellow. The document that generated the search is colored in red and will remain so as one clicks around in the visualization.

Varying node size also reveals information to the user. Federal grants and clinical trials are sized according to their KPIs (key performance indicators).  For federal grants, the larger the circle, the larger their amount of funding.  For clinical trials, the larger the circle, the larger the enrollment size.

Shared rare words can be viewed in the word cloud located in the bottom, right-hand corner of the screen.  Clicking on a word in the word cloud highlights other documents in the visualization that share that word.  Hovering over the small circular "i" icon (for "information"), located in the upper right hand corner of the visualization, will remind the user of these display conventions.

As the user clicks on the document nodes in Omnity’s timeline visualization, the *info panel* on the right-hand side will update with information about the selected document.  The document’s title shows in orange at the top of the *info panel*.  Clicking that title will open the original document - e.g. patent (uspto.gov), clinical trial (clinicaltrials.gov) , or federal grant (grants.gov).

Furthermore, the user is not limited to just the original search.  The first search results are based on the patent the user clicked  from ***NIH / USPTO Cancer Network Miner***.  Double clicking on any node in the timeline visualization will generate a new search based on that document, and the visualization will update accordingly, i.e. with search results related to the new document.  This new search document is shown in purple.  When the original document is in the search results, its node will continue to be red.

The timeline visualization incorporates time, organization, and semantic relatedness, as shown through the rare shared words displayed in the *word cloud*. The visualization assists the user in revealing otherwise hidden conceptual connections shared between patents, research grants, and clinical trials.

You can return to the ***NIH / USPTO Cancer Network Miner*** by selecting its tab in your web browser.

1. NIH Grant funding data was obtained by creating a list of unique grants from the “NIH\_Federal\_Grant\_Number” field in the Cancer Data12A file, then querying grant information from the NIH RePORTER website (https://projectreporter.nih.gov/reporter.cfm). [↑](#footnote-ref-1)
2. We created a list of patent numbers from the Cancer Data12A data set, which was fed to a web scraping program that opened the USPTO web page for each patent and extracted the abstract from the HTML code (for example, http://patft1.uspto.gov//netacgi/nph-Parser?Sect1=PTO1&Sect2=HITOFF&d=PALL&p=1&u=%2Fnetahtml%2FPTO%2Fsrchnum.htm&r=1&f=G&l=50&s1=8529951 .PN.&OS=PN/8529951 &RS=PN/8529951 for patent # 8529951). [↑](#footnote-ref-2)