Hypergraph visualization and enrichment statistics: how the EGAN paradigm facilitates organic discovery from Big Data

Jesse Paquette and Taku Tokuyasu

1University of California, San Francisco
jessepaquette@gmail.com

ABSTRACT

The EGAN software is a functional implementation of a simple yet powerful paradigm for exploration of large empirical data sets downstream from computational analysis. By focusing on systems-level analysis via enrichment statistics, EGAN enables a human domain expert to transform high-throughput analysis results into hypergraph visualizations: concept maps that leverage the expert’s semantic understanding of metadata and relationships to produce insight.

visualization, enrichment, metadata, big data, organic intelligence, data integration, multivariate statistics, cloud computing

1. BACKGROUND

Sets

Humans organize things in their environment into semantically meaningful sets. Natural language is a great example: an adjective is an annotation label that can be associated with one or more nouns; every noun X associated with adjective Y is an element of set Y. Nouns can also be sets themselves; the phrase “X is a Z” can be transformed into the logical concept “noun X is an element of set Z”. These natural language principles reflect an aspect of human cognition that has persisted across millennia. In today’s computational age, this process of entity-to-set association has exploded into a universe of data.

Consider a social network where entities are people in the network. Potential person-sets could be: hometown, current location, alma mater, current employer, first name, last name, movies or other media people like, product advertisements people have clicked on, games people play, posts people have commented on, hashtags people have used, and social contacts of one or more people; just to scratch the surface. The more broadly one expands the definition of person-sets, the richer the data describing each person-entity. This same concept applies to genomic research – where tens of thousands of genes have been annotated with tens of thousands of Gene Ontology terms hundreds of thousands of times1, to media libraries – where media items can be grouped and categorized (e.g. this paper has metadata tags as well as n-grams), to retail products, to companies listed on a stock exchange, to fantasy football results, etc.

The actual data warehouses that store all this information may be arranged into loose, almost unstructured schemata or complex thousand-table relational database systems. The paradigm explored in this paper transforms all these models into a simple schema: 1) there are entities that are the focus of domain-specific research (e.g. people, genes, media items), 2) there are potential network connections between those entities (e.g. personal relationships, protein-protein interactions, nearest-neighbor media, hyperlinks), and 3) there are sets of entities, partitioned into set-categories (e.g. San Francisco, California as a set of people-entities in the location set-category, and UCSF as a set of people-entities in the alma mater set-category; there may also exist a different set UCSF in the employer set-category).

This schema is essentially a simple form of topic map2, where entities in this paper are equivalent to subjects in topic map nomenclature, and sets are equivalent to topics. The difference is that our schema does not attempt to represent relationships between different sets/topics. For example, a topic map might explicitly model that the employer set/topic UCSF has an is-in relationship with the location set/topic San Francisco, California. The EGAN3 schema takes advantage of the fact that a human analyst, a domain expert, will be interpreting the information; that expert will provide their own complex mental map about how sets/topics are semantically related. Keeping the schema simple allows for more metadata and sets to be included and it keeps the workflow simple for the user.
Empirical data

All of the entity types that can be grouped into sets as described above are monitored and researched via collection of empirical data; and the information is flowing in from all directions: numbers, change rates, clicks, purchases, scores, votes, surveys, ratings and more. The analytics/prediction industry is evolving right along with the empirical input stream with algorithms for clustering, classification and prediction; all of which can parallelized to an array of cloud-based processors in order to find that needle in the haystack as quickly as possible. Many of these algorithms work on matrices of data, where each column in the matrix represents an entity and each row represents a variable that can be measured for each entity.

There are two problems with the current analytics paradigm for large empirical data sets (beyond storage and parallelization issues): noise/sparseness and single needle focus. The issue of noise/sparseness depends on the quality of the data collection process and the consistency and frequency of the variables one is hoping to analyze. All empirical data sets have some degree of noise; even a little noise in the data begs the question: how do we know that the best-candidate needle found by an algorithm is really the needle? What about the second-best candidate? What about the tenth-best? What about the thousandth-best? Depending on the number of entities and the strength of the confidence values produced, one might want to consider many more candidates than just the top-hit. It may be easy for an analyst to manually investigate one candidate, but what about the top hundred candidates?

The other problem is single needle focus. Very often an investigator is not interested in simply finding just one needle in the haystack; instead they want to learn how the best candidates from the analysis, or a subset of entities that cluster together, are related to each other, and what those relationships indicate about the environment measured in the experiment. The hypothesis that drives this type of experiment is a systems-hypothesis: no individual entity in the environment is as important as different systems (sets!) of entities. Systems-driven knowledge discovery can identify important trends like social trends, purchasing behavior of consumers, hidden drivers of markets, communication flow in networks and novel biological processes in disease.

This paper presents a systems-approach to analytics that follows downstream of high-throughput multivariate analysis. We do not attempt to outperform and replace your favorite algorithm; rather we intend to provide a paradigm (and software tool implementing that paradigm, EGAN) that facilitates knowledge discovery by integrating the results of your favorite algorithm with a knowledge base of entities and semantically meaningful sets. The secret ingredient in EGAN is not an algorithm, but rather visual representation of this data in a format that sparks the power of human imagination.

2. THE EGAN PARADIGM

Exploratory Gene Association Networks\(^2\) (EGAN) was invented to help bench biologists interpret gene lists (i.e. top-hits) that were produced by different types of high-throughput omics (i.e. genomics, transcriptomics, proteomics, etc.) experiments. There is commonly a lot of noise in omics data, so biologists often struggle through their gene lists for a best candidate or candidates; there are many false positives lurking that will fail to repeat in follow-up experiments. Biologists are also interested in discovering the systems (gene sets, pathways) that will shed light on the biology behind each experiment; implication of specific cellular systems can suggest interesting follow-up experiments and can insure against the risk of follow-up failure with individual genes. Perhaps gene X won’t produce the same result in a follow-up experiment, but system S, of which gene X is an element, will.

Let’s say we have an analysis result, from an experiment with hundreds of thousands of entities, and we have a statistic and p-value calculated for each entity. Sure, we can focus individually on the entity with the top score, but what can the analysis result tell us about the systems involved?

Enrichment statistics

The first step in the EGAN process is to calculate enrichment for each set in the knowledge base from a given analysis result. A set S is enriched if a high degree of the top-scoring entities from an experiment are elements of S, beyond what would be expected by random chance. For example, Caucasian males are an enriched set in the U.S. Congress, compared to the U.S. population as a whole. There are two method types that EGAN uses to calculate enrichment scores: set-based enrichment and global enrichment.

Set-based enrichment is dependent on the selection of a discrete set of top-hit entities, or by taking a discrete cluster, from an analysis result. Often times this is done by employing a statistic (often p-value) cutoff; entities with scores better than the cutoff are subsequently declared to be top-hits, the rest are not. The method then calculates a statistic, commonly a Fisher’s Exact Test\(^4\), to determine for each set in the knowledge base if it is over-represented in the set of top-hits. It is important to note that many cutoff values are selected ad hoc, which
introduces artifacts in the enrichment scores; EGAN accounts for this issue by allowing users to dynamically adjust cutoff values to produce different compositions of the top-hit list.

Global enrichment does not require an explicit declaration of top-hits; it simply considers the ranking of all entities in the analysis result by a given statistic. Those sets with significantly higher ranks compared to a random distribution (usually done by Kolmogorov-Smirnov test) are considered to be enriched. Global enrichment does not suffer from ad hoc introduction of artifacts, but it does give equal weight to entities that do not have particularly significant confidence statistics.

EGAN employs both methods for calculating enrichment, based on the belief that they each provide a different, but valuable perspective on an experiment; scores produced by the two methods can be analyzed separately or taken together.

Hypergraph visualization

A commonly asked question upon calculation of enrichment scores for sets is “what are the top-hit entities in set S that caused it to receive a significant enrichment score?” EGAN was initially built to answer this exact question; it does this via hypergraph visualization.

First: a quick primer on graph visualization. A typical graph contains two types of visual components, nodes and edges. Nodes are commonly rendered as simple shapes with text labels. Edges are rendered as lines that connect two nodes. In the EGAN schema, entities are manifested as nodes and edges represent relationships between entities.

Graphs become hypergraphs when there is additional information in a graph that groups entity nodes together into sets. Complex hypergraphs with many entities and many sets are not easily visualized, however; here is a quick overview of the different hypergraph visualization method (HVM) options.

The traditional HVM is the Venn diagram [Figure 1A], which represents each set as a curve enveloping all entities in the set. The Venn diagram displays set-set overlap nicely; but only when there are less than four sets displayed. Requiring three or less sets is a critical limitation for Venn diagrams.

The clique HVM represents a set as multiple edges that fully-connect all entities in the set [Figure 1B]; but the number of edges required scales poorly (n-squared) with set size.

The node-coloring HVM, gives all entities in a set the same color or shape [Figure 1C]; but it has three important limitations: 1) entities associated with more than one set must be divided into differently-colored sections, 2) the number of displayable sets is limited to a set of distinguishable colors, and 3) finding “all entities that are elements of set S” requires a visual scan of every entity on the graph.

The EGAN application renders entities in its graph visualization as nodes and sets of entities as association nodes (alternatively metanodes) [Figure 1D]. Association nodes are given different shapes and colors to be distinguishable from entity nodes. The association node HVM employed by EGAN allows the user to directly locate a set on the graph as an association node and then visually follow edges from that association node to the nodes that represent entity members of that set. It scales well with respect to the size of each set as well as the number of sets shown on the hypergraph. Association nodes can also be taken into consideration as regular nodes by graph layout algorithms, which are then able to arrange entity nodes in close proximity to the sets in which they are elements, producing a visualization that represents a data-driven concept map.

We have attempted to produce in EGAN a visualization method that most effectively maps pertinent information from an analysis to semantic concepts in mind of the domain expert. When this method is successful, the combination of the rendered information with the expert’s understanding of metadata terms and experiment design evokes high quality conclusions and new ideas.

Workflow

EGAN does not make automatic presumptions about which entities and sets are most interesting with respect to the experimental data; it defers to the expertise of the domain expert. There are a wide array of user interface features in EGAN that help an expert to investigate top-hit entities, calculate enrichment statistics and ultimately arrange entities with association nodes in a hypergraph visualization. EGAN also makes heavy use of Cytoscape libraries, which provide many useful features like graph rendering, user interactivity and layout algorithms.

A typical EGAN workflow starts with the domain expert loading one or more analysis result files (each consisting of a statistic and p-value for all entities measured). The expert then selects significant entities based on statistics and/or set membership (perhaps the expert has loaded clusters from the experiment as custom sets). Next,
the expert invokes set-based enrichment statistics to calculate a measure of over-representation for all sets of entities in the significant set; and optionally global enrichment statistics. The expert employs a combination of enrichment statistics and semantic meaning to select the most interesting sets. Finally, the significant entities and the most interesting sets are displayed as a hypergraph and arranged via automatic layout algorithm and manual manipulation. EGAN provides functionality for the expert to export the visualization to PDF and/or save the analysis for future use.

3. EXAMPLES

Omics

Data from Mirzoeva et al. 9 were analyzed for differential gene expression and DNA copy number between luminal- and basal-type breast cancer cell lines. Figures 2-6 show the progression of the workflow from selecting interesting gene entities via experiment analysis results [Figure 2] [Figure 3], to visualization of those genes on a graph [Figure 4], to enrichment statistics [Figure 5] to visualization of interesting genes and selected enriched sets on the hypergraph [Figure 6]. A biologist can interpret the hypergraph [Figure 6] on multiple levels; to suggest a few: 1) the experiment recapitulates results from similar experiments as evidenced by the enriched MSigDB5 and GeneSigDB10 set nodes, 2) the enriched cytoband regions and their associated genes could be investigated as biomarkers, 3) the implicated pathways from KEGG11 and Reactome12 could be investigated for their relationship to basal-luminal differences, 4) the enriched Medical Subject Heading13 (MeSH) term Papain and the drugbank.ca14 node Clobetasol suggest therapies, and 5) the gene ANXA1 which stands out among the rest with the highest differential expression in basal-type cell lines presents an interesting candidate for follow-up.

Clinical biomarkers

Gene expression for 1143 Breast cancer tumor samples from Gatza et al.15 were analyzed via Principal Components Analysis (PCA); gene expression data for 56 breast cancer cell lines16 were projected into the same principal component space. [Figure 7] shows a scatterplot of PC1 vs. PC2 for the cell lines. Sets displayed on the hypergraph represent the previously characterized subtype of the cell lines. A biologist can interpret this graph as a confirmation that the 56 cell lines represent an adequate subset of breast cancer tumor types; conclusions about breast cancer based on investigating these 56 cell lines will be more useful than if the cell lines only represented one or two subtypes of breast cancer. If the latter were the case, then the cell line sample nodes would not be so scattered on both axes, and they would not be organized in two-dimensional space in a similar pattern that matches their previously characterized subtype. Additional sets could be used to further characterize and investigate these cell lines: mutation status of key genes (p53, PTEN, AKT, etc.), protein expression (ERBB2, EGFR, ER) and other biomarkers/clinical features.

ROI – web log analysis

Apache logs from the web server at http://akt.ucsf.edu/EGAN/ were filtered (image and other requests removed) and parsed into EGAN. [Figure 8] shows a hypergraph of all entity nodes (each is a single HTTP request) for one day, 2011-01-19, with set nodes representing GeoIP17 (i.e. a best guess at the geographical location of the visitor’s IP address) and HTTP referer (i.e. the site from which the visitor came). This visualization is useful for web-based return on investment (ROI) analysis; most visitors come to EGAN from the Gene Ontology website1, others come from the original EGAN paper at the journal Bioinformatics3, and occasionally visitors are discovered to have followed a previous EGAN user’s blog post or tweet.

Media/retail analysis

A common use-case in retail analysis attempts to answer the question: “given the following subset of products we have identified by statistical analysis (e.g. clustering, top-ranked items in a geographic region, etc.), what can we learn about overall purchasing trends?” [Figure 9] scratches this surface by showing the top 30 DVDs from Amazon.com in 200618. DVD entity nodes are connected to sets representing actors, directors, producers, and genre. One can see how the hypergraph in this case serves a similar purpose to a Flikr tag cloud19; in this case the tags/sets are selected for display by set-based enrichment statistics.

Social network analysis

[Figure 10] shows a collection of tweets recorded during the 2010 State of the Union Address20. Tweet entities are shown with text matching “barack”, “obama”, “president”, “union” and “SOTU” and for users with
locations matching searches for “san francisco”, “new york”, “chicago”, “texas” and “atlanta”. For each location-specific set of tweets, enrichment scores for sets representing words used in each tweet were calculated. The top enriched word sets for each location were made visible on the hypergraph. This example uses the tweets themselves as entities, but it would also be interesting to focus on users as entities and see them connected directly by following/friend/in-reply-to relationships and see them associated by metadata sets (data not available); this would also be applicable to Facebook and other social networks.

**Unidentified Flying Objects**

[Figure 11] shows a network of North American UFO sightings, associated with enriched state sets from which they were reported and enriched word sets from the sighting reports. Sighting entities are shown for only one day, 1999-11-16, the day of a significant Leonid meteor shower, especially significant in the Midwestern United States. It is interesting to note how the word sets do not seem to be specific to individual states or even neighboring states, which indicates that the phenomena were rather uniformly perceived.

**4. DISCUSSION**

There exist software tools that calculate enrichment statistics for sets of entities, and there exist software tools that display entities on a graph; even some that represent sets as association nodes. The novelty of EGAN lies in the combination of these different types of functionality, and the analysis paradigm that flows from high-throughput analysis results to user-generated hypergraphs showing significant entities and enriched sets. EGAN additionally provides options for deeper investigation of interesting entities, sets and relationships; each of those items can be linked to knowledge base websites and literature. In the omics realm, the ability to link to Entrez Gene (for each gene entity), the Gene Ontology website (for each Gene Ontology gene set), pathway databases (for each pathway gene set) and PubMed (a literature repository) is extremely powerful.

EGAN does not model or display relationships between two entity sets, as a topic map commonly would. Gene Ontology terms, for example, are actually organized in a tree-like hierarchy. This would be an interesting feature to explore with respect to modeling, visualization and enrichment algorithms.

The relationships rendered as edges in EGAN (entity-entity relationships and entity-set associations) are discrete: a relationship exists or it does not. Many relationships in the knowledge universe are actually fuzzy; so it may be better to allow a confidence score for each association. EGAN could then render the edges with different transparency or provide the user with a slider widget to adjust an edge confidence score cutoff. At present, the user must employ a cutoff for this information before loading the data into EGAN: not very dynamic.

Additionally, the relationships between entities are strictly undirected edges in EGAN at present. This model works just fine for protein-protein interactions between gene entities or friendships between person-entities, but it fails to represent other types of relationships, like when “person X called person Y” (instead EGAN can only model “there was a phone call between persons X and Y”) or when media item-entity A has a hyperlink to media item-entity B.

Clearly not all data sets will be easily flattened into the EGAN data model (entities, relationships, sets); and it is likely that EGAN, which is a generic, yet “skinnable”, analysis tool, may not be valued as highly as domain-specific analysis tools with complex domain-specific data models. The established value of the EGAN paradigm for omics analysis, however, along with the proofs-of-concept for other entity-domains shown above, indicate that the simple EGAN schema provides value for many domains of data that could otherwise be modeled much more elaborately.

Quality of metadata is of critical importance for successful EGAN use. In the biology realm, the network of entity-set associations is biased by the state of current knowledge models about the subject; one task of the biologist is to infer which edges in the network are most likely to be missing. EGAN has been built to accommodate as many diverse sources of metadata as possible in order to provide its users with the most comprehensive perspective.

EGAN is currently stable and available from http://akt.ucsf.edu/EGAN/. The website and all documentation are currently tailored to use of EGAN in biological omics research; but the application is ready to be skinned for other research domains. EGAN runs on a client computer via Java and the amount of data it can handle is dependent on the available memory on the client computer. For the omics case, tens of thousands of genes, a hundred thousand gene sets and a few million edges can be loaded in about 1.5GB of RAM. 64-bit processing architectures allow for greater RAM consumption by any single application, so they are preferred for EGAN use.

SPIE-IS&T/ Vol. 7865  78650E-5
Products of EGAN use (conclusions, new hypotheses, images) are largely subjective and should be understood as such. Most ideas facilitated by EGAN require follow-up experiments for confirmation.

ACKNOWLEDGEMENTS

The authors wish to acknowledge Ajay Jain, Adam Olshen, Stephan Gysin, Antoine Snijders, Ingrid Revet, Sook Wah Yee, Mauting Lin, Donna Albertson, Joe Gray, Debo Das, Eric Collisson, Robert Blelloch, Rob Judson, Mike Nehil and Ben Kopman for their advice, assistance and support. Funding for EGAN development was provided by the National Institutes of Health (grant number P30 CA92103)

REFERENCES


Figure 1. Shown are four different hypergraph visualization methods (HVMs) for the same hypergraph consisting of 9 entity nodes (labeled 1-9) and 4 association nodes (labeled W-Z). 1A shows the Venn diagram; note the pseudopod required to show the overlap between sets X and Z. 1B shows the clique method; note how set X, which contains six entity nodes, requires fifteen edges to represent the set. 1C shows the node-coloring method; note how node four, a member of 4 sets, is forced to divide into four sections in order to display four colors. Also, note how one must scan every node in order to find out which entities are members of set X (teal-colored). 1D shows the association node HVM; the method implemented in EGAN. Note how the entities and the sets form a concept map that clearly and succinctly portrays the associations between entities and sets.
Figure 2. A spreadsheet-like table showing experimental analysis statistics loaded for gene entities. There are two experiments loaded; each experiment has three columns (statistic, sign of statistic and p-value). For the first experiment, genes with positive statistics and small p-values are found to have higher mRNA expression in luminal-type breast cancer cell lines. Genes with negative statistics and small p-values are found to have higher mRNA expression in basal-type breast cancer cell lines. Green indicates positive statistics, blue indicates negative statistics. The second experiment performed the equivalent analysis on array-CGH, which measures DNA copy number instead of mRNA expression. The table is sorted by the p-value column for the mRNA expression analysis.
Figure 3. A dialog which enables the EGAN user to set an empirical cutoff for genes using multiple experiment results.
Figure 4. A graph showing the 36 gene entities selected by the dialog criteria in Figure 3 and their direct relationships. Intensity of blue indicates intensity of differential mRNA up-regulation in basal-type breast cancer cell lines as compared to luminal-type. Width of border is proportional to the log(inverse) of each gene’s p-value. Pink edges indicate protein-protein interactions, gray edges indicate chromosomal adjacency and orange edges indicate that the connected gene nodes are both mentioned in one or more of the same journal articles.
Figure 5. A spreadsheet-like table showing enrichment statistics calculated for MSigDB C2: chemical and genetic perturbations set nodes, representing gene lists published in previous omics experiments, given the chosen 36 significant genes in Figure 4. The most enriched gene set (lowest score, shown at the top) is “Genes whose expression is consistently negatively correlated with estrogen receptor status in breast cancer - higher expression is associated with ER-negative tumors”. This is consistent with our analysis, because basal-type breast cancers are generally ER-negative.
Figure 6. Enriched set nodes are included on the network. Numeric values indicate enrichment scores (not adjusted for multiple hypothesis testing).
Figure 7. A network showing entity nodes (cell lines) arranged in a scatterplot where the x axis represents PC1 from the projection of each cell line’s mRNA expression data into principal component space derived from the tumor samples in Gatza et al. (2010). The y axis represents PC2. Set nodes are shown that represent the previously-known subtype for each cell line. Node positions have been slightly adjusted for label readability.
Figure 8. A hypergraph showing entity nodes (HTTP requests) for http://akt.ucsf.edu/EGAN on 2011-01-19, organized by IP address (gray edges connect visit nodes from the same IP), geographic location and referer (the website that brought the user to the EGAN website). Clusters of fully connected apache visits indicate use of the EGAN software. Some HTTP requests have been filtered out from this set (e.g. image requests).
Figure 9. A hypergraph showing entity nodes (30 top-selling DVDs at Amazon.com in 2006), associated with enriched set nodes (Amazon.com category metadata). Figure prepared by Ben Kopman.
Figure 10. A hypergraph showing entity nodes (tweets concerning the 2010 State of the Union Address), associated with selected Geographic Region sets and selected enriched words for each region. Figure prepared by Ben Kopman.
Figure 11. A hypergraph showing entity nodes (all recorded UFO sightings on 1999-11-16) with selected enriched State and Word set nodes. Figure prepared by Ben Kopman.