Eurographics Conference on Visualization (EuroVis) 2015
H. Carr, K.-L. Ma, and G. Santucci
(Guest Editors)

Volume 34 (2015), Number 3

Refinery: Visual Exploration of Large, Heterogeneous Networks through Associative Browsing

S. Kairam[1], N. H. Riche[2], S. Drucker[2], R. Fernandez[2], and J. Heer[3]

[1]Computer Science Department, Stanford University
[2]Microsoft Research
[3]Computer Science & Engineering, University of Washington

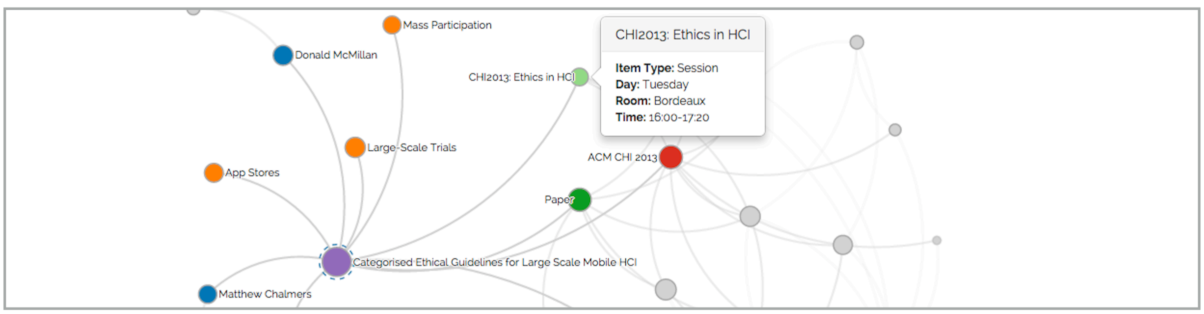

Figure 1: *Refinery allows exploration of large, heterogeneous networks by visualizing subgraphs of items relevant to user queries. Here, querying for a paper in a publication network has led this user to a conference session of possible interest.*

Abstract

Browsing is a fundamental aspect of exploratory information-seeking. ***Associative browsing*** *represents a common and intuitive set of exploratory strategies in which users step iteratively from familiar to novel bits of information. In this paper, we examine associative browsing as a strategy for bottom-up exploration of large, heterogeneous networks. We present Refinery, an interactive visualization system informed by guidelines for associative browsing drawn from literature on exploratory information-seeking. These guidelines motivate Refinery's query model, which allows users to simply and expressively construct queries using heterogeneous sets of nodes. This system computes degree-of-interest scores for associated content using a fast, random-walk algorithm. Refinery visualizes query nodes within a subgraph of results, providing explanatory context, facilitating serendipitous discovery, and stimulating continued exploration. A study of 12 academic researchers using Refinery to browse publication data demonstrates how the system enables discovery of valuable new content, even within existing areas of expertise.*

Categories and Subject Descriptors (according to ACM CCS): H.5.2 [Information Interfaces and Presentation]: User Interfaces—H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—

1. Introduction

Navigating electronic collections often requires a variety of strategies, which have been classified broadly into two categories, *analytical* and *browsing* [MS88]. While analytical strategies are used for retrieving specific facts, browsing is used more for exploratory information-seeking tasks, defined here as learning about and investigating a knowledge domain via continuous, iterative interaction with available resources [Bel93, Mar06, WKDS06]. Prior research in HCI has found that searchers engaged in exploration often naturally adopt strategies based on *orienteering* [OJ93, TAAK04], navigating towards information goals via small,

iterative steps using cues from the environment. As the term *orienteering* has been used recently to describe navigating towards known information targets [TAAK04, CMF08], we use the more general term *associative browsing* to refer to cases in which the information goal is either a specific topic or general knowledge-gathering. We contrast these strategies with *teleporting*, or using keyword search and other means to navigate immediately to desired pieces of content. In cases where teleporting is impossible or unwieldy, associative browsing offers benefits such as circumventing difficulties in specifying queries [FLGD87,tHPvdW96,TAAK04] and providing explanatory context for results [TAAK04, DCW11].

In this paper, we present Refinery, a visualization system designed to support exploration of large, heterogeneous networks through associative browsing. Despite the varied benefits of associative strategies for exploration, there has been limited work examining how to design visualization systems specifically to support them, especially in the area of network visualization.

We first review literature on exploratory information-seeking, identify guidelines for designing interfaces to support associative browsing, and examine techniques used by several classes of existing systems for instantiating these guidelines. We then describe the design and implementation of Refinery. Finally, we present the results of a study of 12 academic researchers using the system to browse conference publication data. We observe how they use Refinery to explore new research areas and discover novel insights, even within their existing areas of expertise. The primary contributions of this work are as follows:

- We identify design guidelines and strategies for interactive visualization systems to support associative browsing.
- We present Refinery, a system which uniquely instantiates these associative strategies to enable effective bottom-up visual exploration of heterogeneous networks.
- We describe a novel application of random-walk based graph algorithms to the problem of extending *degree-of-interest* (DOI) visualization to heterogeneous networks.

2. Developing Design Goals for Associative Browsing

In this section, we review prior work from several areas relevant to exploratory information-seeking in order to develop a set of high-level guidelines for designing visual interfaces for associative browsing over complex data.

2.1. Background: Exploratory Information-Seeking

For decades, information retrieval focused on matching user queries to documents. As discussed by Marchionini & Shneiderman [MS88], the advent of hypertext collections offered diverse possibilities for navigation, with analytical retrieval tasks complemented by browsing, which is more continuous and iterative in nature. Using the metaphor of *berrypicking*, Bates [Bat89] observed that individuals pick up bits and pieces of information as they navigate through an information space. Belkin [Bel93] highlights how searchers not only accumulate knowledge but also change their perception of the search task through interaction with information in the environment.

Studying individuals searching library collections, O'Day & Jeffries [OJ93] observed how searchers examined results returned by librarians and used these to modify strategies for future search iterations. Teevan et al. [TAAK04] observed similar strategies for users browsing electronic information on their personal computers. Despite the different contexts and goals, searchers in both studies naturally adopted an iterative process of leveraging contextual cues to choose subsequent exploration steps until search goals were achieved.

One observed benefit of adopting associative strategies is in circumventing difficulties in composing queries. The well-known "vocabulary problem" [FLGD87] stems from the fact that searchers often have to choose from many possible search terms for a given target. In addition, difficulty recalling details about the target early in the search often makes keyword search untenable [TAAK04]. For searchers without a clear notion of the target, Bruza [Bru93] observes that they can usually spot relevant information when it appears, offering the interview quote *"I don't know what I'm looking for, but I'll know it when I find it."* ter Hofstede [tHPvdW96] identifies an interactive query formulation loop with three phases — exploration, construction, and feedback — which searchers repeat until achieving the information goal. The Mr. Taggy system [KNPC09] illustrates how tightening this loop through active suggestion of query refinements can facilitate exploration and sensemaking.

When considering users browsing for known information targets, we can relate these empirical observations to psychological theories holding that memory is encoded and retrieved in the form of inter-item associations. Anderson & Pirolli [AP84] describe how observed information triggers associations to content in long-term memory through a process of *spreading activation*. Earlier experimental work by Tulving [TT73] suggests that retrieval of items from memory can be improved by providing particular associations which were present when the information was initially encoded.

In the case of more general exploration, where searchers aim to uncover previously unknown items, we look to literature on serendipitous finding. André et al. [ASTD09] characterize serendipity as the combination of finding unexpected information and the ability to make an intellectual leap to connect that information to what you already know. Marchionini & Shneiderman [MS88] consider such serendipitous finding to be central to the activity of browsing. Dörk et al. [DCW11] advise that serendipity can be encouraged in visualization interfaces by "juxtaposing resources that share unusual facets or relate to one's previous interactions."

2.2. Guidelines for Supporting Associative Browsing

Based on these findings, as well as challenges observed in real-world exploratory tasks, we offer the following guidelines for designing interactive visualization systems to support associative browsing over complex information spaces.

G1. Support navigation across heterogeneous, dynamic collections. Hypertext collections encompass a variety of media types, and an associative browser should enable exploration across both textual and non-textual data (e.g. images or videos) [MS88, Mar04]. As data are frequently combined from disparate sources with different schemas, we require a common language model for flexibly representing bits of content and their relationships [ASTD09]. Computation should ideally be done on the fly in order to accommodate the increasingly dynamic nature of available information resources (e.g. news, blogs, social media) [Mar04].

G2. Balance simplicity and expressivity in representing search intent. In information retrieval settings, keyword-based search can be highly expressive, but the large space of possible choices can lead to roadblocks [FLGD87, Bru93, tHPvdW96]. However, the system must be sufficiently expressive to allow users to choose queries specific enough to focus exploration in areas of interest. Once users have begun exploring, the system should suggest and allow users to easily evaluate possible query refinements [TAAK04].

G3. Refine search intent through continuous dialog with the user. The interface should aim to engage users directly and actively in the information retrieval process [Bel93], favoring a continuous interactive dialog over more formalized turn-taking [Mar06]. To facilitate this dialog, the system should make it easy for users to provide relevance feedback about suggested items [Mar06, KNPC09].

G4. Surface varied contextual cues to support recognition and discovery. Observing that recognition of content improves when it is presented with context matching that in which it was encoded [TT73], the interface should offer a diverse set of contextual cues and connections for known and recommended items. These cues can facilitate recognition of target items when uncovered [Bru93, TAAK04] or serendipitous finding of useful, novel items [ASTD09, DCW11].

2.3. Related Systems

We examine here several classes of systems related to exploratory information-seeking. For each class, we highlight an example system to illustrate techniques for instantiating these guidelines, summarizing our observations in Table 1.

Faceted and Cluster-Based Browsing. This class of interfaces allow users to leverage item metadata to iteratively explore collections. *Faceted browsing* allows users to specify filters using metadata to find subsets of items sharing specific desired characteristics. A study of the Flamenco system [YSLH03] illustrated how metadata could help users browse large image collections more easily than keyword search alone. *Clustering* achieves a similar goal by using metadata to group items with similar properties. Studying category usage in the Findex system, Käki et al. [K05] observed how clusters helped users to refine queries when initial keyword searches failed. Rodden, et al. [RBSW01] observed the important role that decisions about strategies for categorizing images play in subsequent browsing behavior.

	G1	G2	G3	G4
Faceted/Clustered Browers	$-$	$+$	$+$	$-$
DOI Visualization	$-$	$\diamond$	$+$	$+$
Ostensive Browsers	$-$	$+$	$+$	$\diamond$

Table 1: *For each class of systems, we summarize the extent to which each guideline is supported using $+$, $-$, or $\diamond$ (which indicates mixed support).*

We highlight Flamenco as a well-known example of these techniques. Flamenco combines keyword search with easily selectable facets, providing multiple means of specifying search intent (+**G2**). The system enables users to iteratively select facets and receive suggestions for refinements to navigate easily towards search goals (+**G3**). Flamenco and related systems (e.g. FacetLens [LSR*09] & PivotSlice [ZCCB13]), however, only enable browsing items of a single type and only allow queries based on textual or numerical facets ($-$**G1**). Furthermore, both faceting and clustering hide relationships among items within a visible group or across groups ($-$**G4**), possibly hiding opportunities for exploration and discovery. PivotSlice serves as an exception, displaying relationships between items within such groups.

Degree-of-Interest (DOI) Visualization. DOI visualization techniques, as proposed by Furnas [Fur86], magnify or highlight items of interest along with a subset of items which may provide explanatory context. Techniques to increase the visual saliency of important neighbors have been applied to tree [CN02, HC04] and graph [LPB*06] structures. Both van Ham & Perer [vHP09] and Crnovrsanin et al. [CLYM11] present sophisticated network exploration systems which combine several notions of relevance to compute overall DOI scores and visualize nodes in the graph.

Using van Ham & Perer's system as an example, we see that DOI systems are built around the notion of visualizing contextually relevant information for a particular object of interest (+**G4**). The interactive system they describe allows users to select these contextually relevant items and add them to the current focus, facilitating rapid refinement of the view (+**G3**). It is non-trivial, however, to extend their formulation to heterogeneous networks ($-$**G1**). One problem is defining separate user interest functions for each type of node. Learning to rank documents against images and other types of content for each new dataset would require significant research and refinement. In addition, their approach is open-ended with respect to how queries are specified ($\diamond$ **G2**).

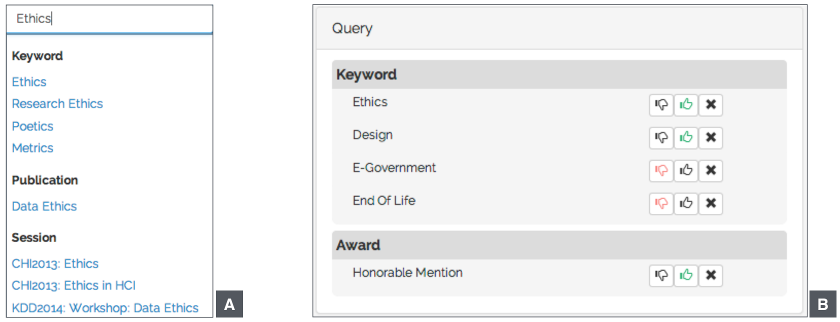

Figure 2: *(a) Free-text search allows users to identify items matching their interest and add them to the query. (b) Items which have been "upvoted" or "downvoted" into the query are grouped together in the Query Panel for easy reference.*

Ostensive Browsers. Several systems for exploratory information-seeking have helped users overcome difficulties in query formulation by attempting to infer the "ostensive relevance" of items [Cam96] directly from the user's interactions with the results. The ViGOR system [HVHJ09] supplements video recommendation using information gained from user-created groupings of relevant results. Apolo, by Chau et al. [CKHF11], similarly allowed users to place results into groups and used belief propagation to suggest additional items of potential relevance. Comparing an ostensive browser to a traditional keyword-search interface for image retrieval, Urban et al. [UJvR06] found that the ostensive browser stimulated significantly more ideas for alternate searches and led to fewer dead ends.

Apolo provides an example of how allowing users to point directly to items to communicate relevance provides a powerful and flexible means of specifying search intent (+**G2**). This interactive loop, with output (results) serving in part as input (queries), allows the user and system to work together to determine relevance (+**G3**). Placing these items in a network visualization, as Apolo does, provides cues about context and relationships, opening avenues for subsequent exploration, but Apolo only shows direct relationships between items of a single type, limiting opportunities for associative browsing (◇ **G4**),. It is not clear how belief propagation might work over heterogeneous networks (−**G1**). In the conference publication dataset used to evaluate Refinery, for instance, high-degree nodes such as *Publication-Type* `Paper` or particular *Conferences* might propagate relevance to large numbers of unrelated items.

3. Introducing Refinery

Based on these design guidelines and insights from existing systems, we have created Refinery, a prototype system for supporting associative browsing over large, heterogeneous information networks. While the approach described generalizes to arbitrary heterogeneous networks, we illustrate the system's features and implementation using academic publication data from the Confer project [BKM14]. This dataset covers 13 conferences from 2012 to 2014 in fields such as HCI, Social Computing, Data Mining, and AI.

3.1. Refinery in Use

Through the following example, we illustrate Refinery's features and how the system can be used for associative finding and browsing: *Mae recalls an interesting talk she attended at a recent conference related to "ethics" in HCI research. She can't recall the title or authors, but she remembers that the paper won an Honorable Mention. She hopes to find that paper, as well as other papers which may be of interest.*

Free-Text Search. Users are presented initially with a search box, inviting them to enter a free-text query. Mae starts by entering "ethics." The search box matches her text against labels of items in the network, pulling up relevant matches (shown in Figure 2(a)). Mae selects the *Keyword* `Ethics`. She uses the free-text box to include a second query item, adding *Award* `Honorable Mention`.

Sidebar. Once an initial keyword has been selected, the main Refinery interface appears. The Sidebar groups query items in the Query Panel at the top for quick reference. Associated items are suggested below in the Facet Panel, grouped by type and sorted by relevance score. Browsing *Keywords*, Mae uses the thumbs-up button to "upvote" *Keyword* `Design`, shifting focus towards this term. She sees the *Keywords* `End of Life` and `E-Government` and recalls that the paper she seeks doesn't address these issues; she "downvotes" these using the thumbs-down button to shift focus away from these topics. Each selection updates her query and the suggested facets being shown. Figure 2(b) shows how the Query Panel represents the current search state.

Graph View. Figure 3 shows the Graph View for this query, visualizing the most relevant items returned by the system along with their connections. Here, Mae browses through items by mousing over each to see connections highlighted, as shown in Figure 1. By clicking on an item, she can see data stored along with that item. She clicks on a *Publication* `Categorised Ethical Guidelines for Large Scale Mobile HCI` and reads the abstract. This isn't the paper she is looking for, but it is very similar, so she adds it to her query in order to attract related content. Doing so causes the *Session* `CHI2013: Ethics in HCI` to appear, which she also adds to the query.

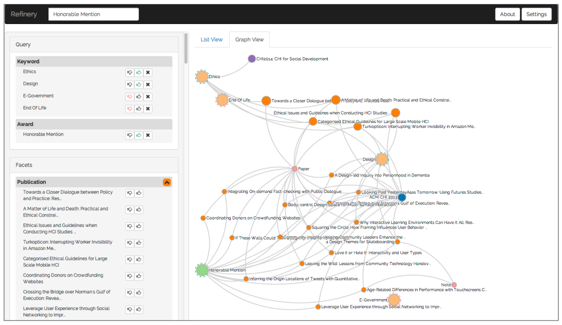

Figure 3: *Refinery's Graph View clearly displays items most relevant to the query along with their relationships.*

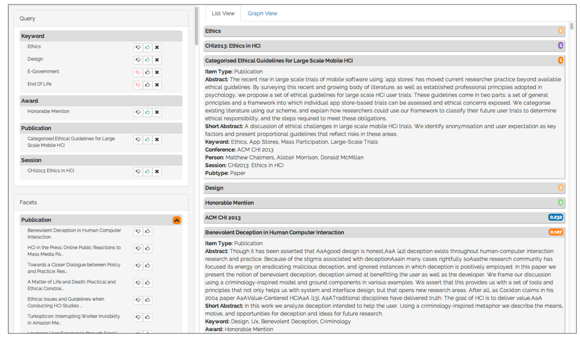

Figure 4: *Refinery's List View ranks items by overall relevance. Users can easily hide or show particular types using toggles in the Sidebar in order to focus their view.*

List View. At this point, Mae feels she is pretty close to finding the paper that she wanted. She switches to the List View (shown in Figure 4), showing all the items returned for her query in a single list, ranked by overall relevance. By clicking the headers for facet groups in the Sidebar, she hides all types except for *Publications*, allowing her to easily scroll through a ranked list of relevant *Publications*. The third item in the list is the paper she had hoped to find, *Publication* `Benevolent Deception in Human Computer Interaction`. Because entries in the List View provide a more complete view of items, she observes that this paper didn't use her original keyword, despite the obvious relevance. She scrolls through this list, finding several other papers closely related to this paper in various ways.

3.2. Modeling the Data

Refinery internally represents information collections as a heterogeneous association network. Node and edge types for the study dataset are given in Tables 2 and 3. Undirected relationships are represented using reciprocal directed edges. Each type of edge has a *weight* attribute, assigned initially based on intuitions about the data (e.g. *Publication-Author* edges represent stronger relationships than *Publication-Conference* edges) and then adjusted through empirical tuning. Our experience with the study dataset found that search results were not highly sensitive to exact choices of edge weight. In the discussion section, we consider how edge weights might be learned or refined based on user interaction with the system.

Each node has unique *key* and *label* attributes used for indexing and display, but nodes and edges are otherwise schema-less, with arbitrary attributes. This representation allows us to represent diverse types of information, including non-textual or multimedia content, and their relationships in a common data structure for browsing and ranking (**G1**). Choosing an ontology for nodes and edges is an important decision, potentially affecting system performance and utility. For instance, rather than separating *Publication-Type* as its own node type, we could have made this an attribute of *Publication* nodes. As a heuristic, any fields with unique values for individual nodes (e.g. title) were treated as attributes. Other categorical attributes were separated into individual nodes (e.g. publication type). Some decisions about how to model a particular dataset may come down to subjective preference or empirical tuning, and a system for projecting tabular data into different network representations, such as Ploceus [LNS11], could prove useful for this task.

3.3. Searching the Association Network

Refinery's interface allows users to specify a query directly using a group of nodes in the network, helping to achieve (**G2**). At a high level, Refinery scores items by "association" to nodes in the query set using an algorithm based on simulated random walks across the network, similar to algorithms such as PageRank [BP98] and Random Walk with Restart (RWR) [TFP06]. Scores are computed separately for each query node and then combined into a single score representing relevance to the entire query set.

For each individual query node, the system simulates multiple random walks starting at this node following outgoing edges. At each step, the walker at some node i will either (1) choose an edge $e_{i\to j}$ from the set of edges e_i outgoing from i with probability proportional to the weight of the edge, or (2) stop with some "halting" probability p_H. The probability that the walker transitions from a node i to a connected node j is given by:

$$p(e_{i\to j}) = \frac{w(e_{i\to j})}{p_H + (1-p_H)\sum_{e\in e_i} w(e)} \quad (1)$$

High values of p_H produce shorter walks, leading to more conservative, locally-relevant results. Low values of p_H produce globally relevant results, similar to those given by PageRank. For each node i in the query set, the score for each other node j in the network is given by the frequency with which simulated walks originating at i ended at j. We assign a score of $n_i = 0$ for any node n that never becomes the destination for a walk from i.

Node Type	**Count**
Conference	13
Session	944
Persona	48
Publication	3200
Keyword	4465
Publication Type	25
Award	2
Person	8203
Affiliation	1822
Total	**18722**

Table 2: *Node types and counts in the Confer dataset.*

Edge Type	**Weight**	**Count**
Publication – Author	10	11252
Publication – Session	5	2780
Publication – Keyword	5	7174
Session – Chair	4	46
Publication – Conference	3	3182
Session – Person	3	225
Publication – Award	2	134
Conference – Session	2	944
Session – Keyword	2	2864
Person – Affiliation	2	6130
Publication – Pub Type	1	1604
Total	**–**	**72670**

Table 3: *Edge types and counts in the Confer dataset.*

For queries composed of multiple nodes, we combine results through simple addition or subtraction of scores. For a query with two nodes, i and j, we add the scores $n_i + n_j$ for each node n in the network. This simple approach intuitively captures content strongly associated with any query item or moderately associated with several. If, instead, the user had "upvoted" i and "downvoted" j, we would subtract the scores $n_i - n_j$. Thus, an item associated with j might still appear highly ranked if its association to i is strong enough.

We choose this approach for a few reasons related to our design guidelines, especially (**G1**). First, it allows us to rank textual and non-textual data of any type in relation to query nodes of any type. In contrast to techniques such s that used in [vHP09], we aren't committed to choosing or pre-computing separate a priori interest or user interest functions for each node type. This is especially useful when considering non-textual nodes. Computing query relevance specifically for videos, for instance, might require advanced image processing techniques.

By simulating random walks at query time rather than pre-computing scores, we can easily handle dynamically updated data, such as social media or event logs. In addition, parameters such as edge weights or halting probability could potentially be tuned by users on-the-fly during search sessions to manipulate results. This approach satisfies our requirements, is suitable for any type of data which can be modeled as a heterogeneous network, and performs as expected for the dataset outlined here. In the discussion, we present some additional datasets to which we have successfully applied this approach.

3.4. User Interaction and Visual Representation

The main interaction in Refinery is adding and removing nodes to and from the query using the "upvote" and "downvote" buttons. This approach, successfully used in the past to foster continuous exploration (as in Mr. Taggy [KNPC09]), affords a simple and lightweight means of providing relevance feedback to refine search intent. Results update immediately after each query update to reflect the current search intent. By enabling a quick interactive loop, we engage users in a continuous, uninterrupted dialog with the system (**G4**).

Refinery provides multiple views on results, offering multiple contexts for retrieving information. Highly ranked items are shown grouped by type in the Sidebar, ordered by score in the List View, and clustered with associated nodes in the Graph View. In the List and Graph Views, higher-scoring nodes are shown with more contextual information than lower-scoring nodes, concentrating the user's attention on items more likely to lead to recognition of desirable content or serendipitous discovery.

As our intent was to foster a novel style of exploration, we adopted the force-directed layout in order to maintain a familiar graphical representation to avoid overwhelming new users. In the future, we might consider constrained layouts for presenting subgraphs, such as extending existing techniques (e.g. [CKHF11,CLYM11]) to heterogeneous networks. In the following subsection, we describe the process which led to the current design; in the Discussion, we consider other potential options for visually presenting results.

3.5. System Development

Refinery was developed through several iterations. An initial pilot with 12 users compared a force-directed layout to a more constrained layout with nodes grouped radially around a circle. The radial layout grouped nodes by type, but the force-directed layout promoted perception of clusters of related nodes of mixed types (better supporting **G4**). Participant feedback indicated that the force-directed layout was 'simpler' to understand.

In the second iteration, we deployed a version publicly for two weeks around the CHI 2014 conference. Conference attendees were invited to use the system to find talks and people of interest. This version contained only the Graph View (no List View). This deployment attracted over 400 unique users, and roughly half of these conducted some exploration. We observed users building diverse query item sets, validating our intuitions that this would be an expressive means

of formulating queries. Requests for more browsable lists of items led to our incorporation of the complementary List View, significantly speeding up result browsing.

Refinery's query engine is implemented in Python, using the `network` library to store and search the association graph. The interface is implemented in HTML/CSS/JS, using the `jQuery` and `d3` [BOH11] libraries for interaction and visualization. `d3`'s force-directed layout is used for the Graph View. As discussed earlier, the system makes available several tunable parameters; to preserve simplicity, these were not exposed to users in the following study. We ran 2,000 walk iterations for each query node, with $p_H = 0.4$ and edge weights as described above.

4. User Study

We conducted a user study with 12 participants to help examine the extent to which Refinery supported our goal of facilitating associative browsing over heterogeneous networks. We specifically recruited academic researchers with expertise in areas covered by the data; we were interested in how Refinery might yield insights extending beyond those uncovered by existing tools, such as Google Scholar.

Participants. We recruited 12 academic and industry researchers (7 Female / 5 Male, Age: $\mu = 27.4$, $\sigma = 3.3$); all had at least some graduate study in the areas of HCI, Information Visualization, or NLP. All participants reported moderate-to-expert familiarity with at least 3 conferences in our dataset. Each experiment session lasted approximately one hour; participants were offered $25 in compensation.

Procedure. Sessions started with a brief introduction and explanation of the study procedure. Participants were given a walkthrough of the dataset and a pre-survey assessing familiarity with the data. The main study task asked participants to explore the data at their leisure for up to 15 minutes; they were informed that they could conclude sooner if they felt that their exploration had drawn to a close. At the end of the session, each participant completed a brief questionnaire about the system (adapted from the evaluation of the Eddi system [BSH*10] by Bernstein et al.) and engaged in a 10-minute interview about their experience.

During the exploration task, we followed a think-aloud protocol, asking subjects to describe their interactions with the system and data. We specifically asked that they report any interesting observations about the data, including errors, surprising patterns or connections, and confirmation or rejections of prior hypotheses or intuitions they may have had. All sessions were run using Google Chrome on a 15-inch MacBook Pro. Study sessions were captured using query logging, screen-capture, and audio-recording.

5. Study Results

We present some behavioral observations and feedback regarding the system in order to assess ways in which Refinery did or did not successfully match our initial design goals. We use codes P1 to P12 throughout when providing examples or quotes from individual participants.

System Usage. All participants used the system actively, including a substantial number of unique nodes as part of their queries over the course of a session (Unique query nodes per user: $\mu = 12.5$, $\sigma = 4.9$). Nodes were combined into diverse queries; participants created a large number of unique query combinations in each session (Unique query sets per users: $\mu = 17.8$, $\sigma = 7.2$). No participant expressed frustration over unintended queries, indicating that Refinery's design supports rapid and expressive query formulation (**G2**).

Most participants created query sets composed of diverse node types (Unique types used per user: $\mu = 3.67$, $\sigma = 1.23$). As we believe Refinery is unique in allowing users to specify queries in terms of heterogeneous node sets in this way, we were encouraged to see this feature used extensively. Every node type, except for *Persona* was used in at least one query by at least one participant. Another method of diversifying queries, "downvoting" nodes, was used at least once by the majority (7/12) of participants in our study.

Subjective Feedback. Based on verbal feedback and our observations of the exploration sessions, participants appeared to enjoy using the system. All (12/12) participants explored for the full 15 minutes, and several asked to continue exploring after the task period had concluded. By design, our study engaged subjects in exploring content information over which they already had some expertise. Despite this high level of familiarity, every (12/12) participant found novel items of interest, based on their own self-report.

One participant, for instance, explored topics relevant to her current research, finding several novel *Publications* which she noted down to review later on her own. Previously unaware of a specific design-related *Keyword* used by authors of these *Publications*. she had missed them earlier when searching using Google Scholar. When discussing how she found them using Refinery, she said, *"It gave me suggestions for things I might not actually have searched for, but were quite related." [P3].*

Another participant searched for ideas to help plan an upcoming research project; he was surprised to find work focused on algorithms being done within the HCI community.

> *It certainly made me more interested in the topic than I was before...I didn't have a genuine deeply-seated interest in it. Now, I think I genuinely do, if only because I see that the way that it is interesting to HCI is the applications of it, and the people whose work...interested me, intrinsically, relates to algorithms and applications thereof. [P12]*

In this case, the user's perspective on his own research was changed by the 'serendipitous' experience of encountering novel information along with associations to content which was familiar and meaningful to him.

Questionnaire Responses. In a post-study questionnaire, participants indicated their level of agreement with several statements using a 7-point Likert scale, from 1 (Strongly Disagree) to 7 (Strongly Agree). Participants rated the system highly in terms of *interestingness* ($\mu = 6.33$, $\sigma = 0.89$), *enjoyability* ($\mu = 5.33$, $\sigma = 1.30$), and *flexibility* ($\mu = 5.25$, $\sigma = 1.36$). Participants disagreed quite a bit about the extent to which they found the system "overwhelming" ($\mu = 3.08$, $\sigma = 2.02$), pointing to possible individual differences in preferences for associative browsing.

5.1. Browsing Strategies.

The study also offered the opportunity to use Refinery as a probe for studying users engaged in exploratory information-seeking within a realistic, but controlled environment. We observed several common low-level strategies which participants mixed and matched while browsing.

Every participant engaged in *refining* (Refinery's main feature), iteratively adding items to the query set in order to focus exploration within a subarea of the data. After building a query set through refining, the majority of participants (10/12) shifted the query focus by *traversing*, removing initial query items until they had navigated to a novel area in the data. Instead of traversing to a new location, participants sometimes *retreated* (5/12), removing newly added items from the query set to return to an earlier view of the data.

In half of the sessions (6/12), we observed at least one situation in which one or more query items served the function of *bridging* between two areas. Adding the "bridge" item(s) prompted participants to remove all other existing query items, add new query items, and continue exploring a new, but related, area of the data. While identifying these strategies was not the focus of the present study, they provide an interesting means of summarizing browsing behavior and could potentially serve as the subject of future research.

5.2. Study Limitations

As Refinery was built for open-ended exploratory tasks, our study aimed to observe users browsing in a realistic and unconstrained setting. Follow-up studies with more specific tasks, such as an insight-based evaluation, might allow for better quantitative comparison of user performance in this system and comparison with others. Participants often verbally compared Refinery to existing search tools backed by more extensive corpora, such as Google Scholar. While the Confer dataset was large enough to facilitate diverse exploration sessions, it was still quite limited compared to those searched by existing tools, making a direct comparative evaluation difficult. While we were encouraged that users discovered novel content even with this limited dataset, it would be helpful in ongoing research to observe users browsing more complete collections to avoid disappointment based on missing content.

In addition, as the interface included many novel elements, some participants required several minutes before feeling that they had 'gotten the hang of it.' The current implementation of Refinery does not preserve node positioning across transitions in the Graph View; adopting techniques for doing so may aid users in tracking objects across multiple queries. Despite efforts to create a realistic task environment, observing participants familiar with the system browsing in self-prompted scenarios may yield different results.

6. Discussion & Future Work

We are encouraged by participants' positive comments, as well as the underlying reasons, which closely aligned with our proposed design guidelines. Several participants explicitly called out the distinction between the associative style of browsing enabled by Refinery and that afforded by more traditional retrieval interfaces, such as Google Scholar.

> *Google Scholar is great when you know what you're looking for...once you know what you're looking for, it's very easy to recognize. But it's not so easy to find things which are ill-defined. [P12]*

> *It wasn't like when you go to Google, and you know exactly what you're searching for, and you just find it. This is more like I'm trying to explore this space and it's a really wide range of things. So being able to put poles where things would gather around the ideas that were really interesting to me was really awesome, and I found articles that I wouldn't have looked at. [P3]*

One participant described how interaction with the data aided her in focusing in exploration, clearly illustrating some of our goals in designing for associative browsing.

> *You don't know what questions you're going to have if you don't know what the layout is...that's how you develop good questions. [P4]*

An important area of future work is exploring how alternative modeling decisions for the association network will impact result quality and exploration strategies. For different types of data, we might consider adding nodes based on computed connections. For text corpora, for instance, we might compute latent topics using LDA [BNJ03] and add nodes for these, connecting documents with similar semantics. For image corpora, we could similarly create nodes to represent computed image characteristics.

While our search algorithm performed admirably for the study data, future work might compare it against related alternatives. An obvious limitation of the current algorithm is that walks from a query node never reach items outside of that node's weakly connected component. We could address this problem by introducing random jumps, as in PageRank [BP98], for instance. The approach of simulating random walks, as we suggest here, also opens up possibilities

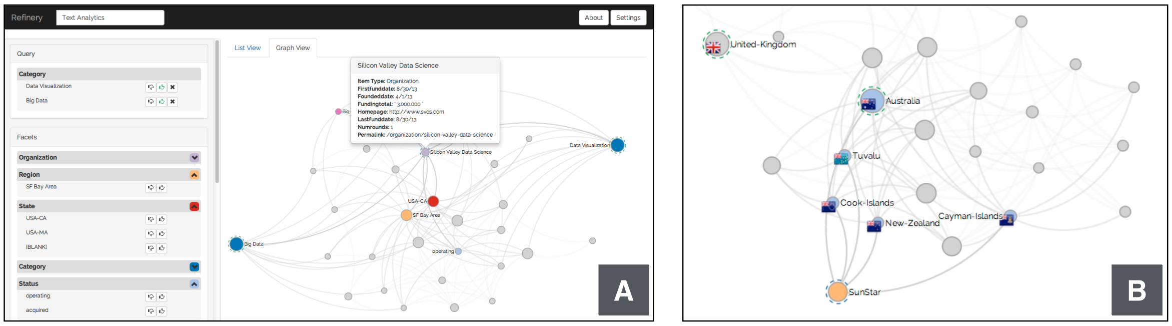

Figure 5: *(a) SF Bay Area startup company data in Crunchbase: looking for content relevant to Big Data and Data Visualization, the user has identified an interesting company and explores further. (b) National flags: viewing content associated with the UK and Australia, the user highlights a node representing "sun or star imagery" revealing that this graphical element is shared by several Commonwealth nations.*

for exploring parallelization as a means for speeding up result computation and scaling to larger datasets. Search could also be improved or adapted by tuning edge weights based on user interaction with the system. For a given search, the system could combine results provided by different parameterizations and update edge weights and parameters based on the results with which users choose to interact.

Our participants expressed enthusiasm for exploring of other types of collections using a similar browsing style. In Figure 5(a), we illustrate how Refinery might be used to explore data about San Francisco Bay Area startup companies from Crunchbase [Cru14]. Here, the user is browsing companies related to Big Data and Data Visualization, discovering a recently-formed company tagged with both of these keywords, and mousing over it for more information.

Figure 5(b) illustrates browsing a dataset of national flags [BL13] which combines textual and non-textual data. Here, flags are linked to nodes representing common graphical elements (e.g. crosses, saltires, or animate figures) and demographic features (e.g. language, continent, religion). Here, after querying for the United Kingdom and Australia, the user mouses over a node representing sun and star imagery, highlighting a graphical feature common to the flags of several Commonwealth countries.

We note here that the flag nodes are simply images, retrieved using only the network structure, illustrating how our approach can adapt flexibly to multimedia content. These nodes could just as easily represent audio files or video files without the need to compute content-specific user interest functions. While our ranking approach generalizes easily to such multimedia data, we might consider alternate visual presentation approaches to aid in navigating such collections.

Media collections, product databases, and news repositories are just some of the areas ripe for next-generation visual interfaces to support associative browsing. We are eager to continue iterating on all of these aspects of Refinery in order to aid users in applying associative browsing strategies for exploratory tasks in these various domains.

7. Conclusion

In this paper, we have introduced Refinery, an interactive visualization system supporting bottom-up exploration of large, heterogeneous networks through associative browsing. We have contributed guidelines for designing visualization systems to support associative browsing which have informed Refinery's design. Through a novel application of random-walk based algorithms to the domain of degree-of-interest visualization, we have contributed a simple, scalable technique for DOI scoring over heterogeneous networks.

Refinery's interface allows users to specify the 'frontier' of their knowledge using collections of nodes of various types, and the system visualizes results to effectively display context and connections which spur insights and further exploration. Results and feedback from a user study with 12 researchers browsing within their areas of study demonstrates that Refinery can aid in exploratory information-seeking, even in cases where users had considerable expertise in the knowledge domain being explored. We hope that the design guidelines and techniques considered here help to inform future visualization systems aimed at supporting associative browsing as a strategy for exploration.

8. Acknowledgements

This research was supported in part by the Defense Advanced Research Projects Agency (DARPA) XDATA program. We would like to thank Michael Bernstein, Stuart Card, Susan Dumais, and Marian Dörk for their helpful comments and suggestions throughout this research project.

References

[AP84] ANDERSON J. R., PIROLLI P. L.: Spread of Activation. *Journal of Experimental Psychology 10*, 4 (1984), 791–798. 2

[ASTD09] ANDRÉ P., SCHRAEFEL M., TEEVAN J., DUMAIS S. T.: Discovery Is Never by Chance : Designing for (Un) Serendipity. In *Proc. ACM C&C* (2009), pp. 305–314. 2, 3

[Bat89] BATES M. J.: The design of browsing and berrypicking techniques for the online search interface. *Online Review 13* (1989), 407–424. 2

[Bel93] BELKIN N. J.: Interaction with Texts : Information Retrieval as Information-Seeking Behavior. In *Proc. of Info. Retr.* (1993), pp. 55–66. 1, 2, 3

[BKM14] BHARDWAJ A., KARGER D., MADDEN S.: Confer Project, 2014. 4

[BL13] BACHE K., LICHMAN M.: UCI Machine Learning Repository, 2013. 9

[BNJ03] BLEI D. M., NG A. Y., JORDAN M. I.: Latent Dirichlet Allocation. *Journal of Machine Learning Research 3* (2003), 993–1022. 8

[BOH11] BOSTOCK M., OGIEVETSKY V., HEER J.: D3: Data-Driven Documents. *IEEE transactions on visualization and computer graphics 17*, 12 (2011), 2301–9. 7

[BP98] BRIN S., PAGE L.: The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems 30* (1998), 107–117. 5, 8

[Bru93] BRUZA P. D.: *Stratified Information Disclosure: A Synthesis between Information Retrieval and Hypermedia*. PhD thesis, University of Nijmegen, The Netherlands, 1993. 2, 3

[BSH*10] BERNSTEIN M. S., SUH B., HONG L., CHEN J., KAIRAM S., CHI E. H.: Eddi : Interactive Topic-based Browsing of Social Status Streams. In *Proc. ACM UIST* (2010), UIST '10, ACM, pp. 303–312. 7

[Cam96] CAMPBELL I.: *The ostensive model of developing information needs*. PhD thesis, University of Glasgow, Scotland, 1996. 4

[CKHF11] CHAU D. H., KITTUR A., HONG J. I., FALOUTSOS C.: Apolo : Making Sense of Large Network Data by Combining Rich User Interaction and Machine Learning. In *Proc. ACM CHI* (2011), pp. 167–176. 4, 6

[CLYM11] CRNOVRSANIN T., LIAO I., YINGCAI W., MA K.-L.: Visual recommendations for network navigation. In *EuroVis* (2011). 3, 6

[CMF08] CHAU D. H., MYERS B., FAULRING A.: What to do when search fails: finding information by association. *Proc. ACM CHI* (2008), 999–1008. 2

[CN02] CARD S. K., NATION D.: Degree-of-Interest Trees: A Component of an Attention-Reactive Interface. In *Proc.ACM AVI* (2002). 3

[Cru14] CRUNCHBASE: Crunchbase, 2014. 9

[DCW11] DÖRK M., CARPENDALE S., WILLIAMSON C.: The Information Flaneur : A Fresh Look at Information Seeking. In *Proc. ACM CHI* (2011), pp. 1215–1224. 2, 3

[FLGD87] FURNAS G. W., LANDAUER T. K., GOMEZ L. M., DUMAIS S. T.: The Vocabulary Problem in Human-System Communication. *CACM 30*, 11 (1987), 964–971. 2, 3

[Fur86] FURNAS G. W.: Generalized fisheye views. *ACM SIGCHI Bulletin 17*, 4 (1986), 16–23. 3

[HC04] HEER J. M., CARD S. K.: DOITrees Revisisted: Scalable Space-Constrained Visualization of Hierarchical Data. In *Proc. ACM AVI* (2004), pp. 421–424. 3

[HVHJ09] HALVEY M., VALLET D., HANNAH D., JOSE J. M.: ViGOR: A Grouping Oriented Interface for Search and Retrieval in Video Libraries. In *Proc. ACM JCDL* (2009), pp. 87–96. 4

[KÖ5] KÄKI M.: Findex: search result categories help users when document ranking fails. *Proc. ACM CHI* (2005), 131–140. 3

[KNPC09] KAMMERER Y., NAIRN R., PIROLLI P., CHI E. H.: Signpost from the Masses : Learning Effects in an Exploratory Social Tag Search Browser. In *Proc. ACM CHI* (2009), pp. 625–634. 2, 3, 6

[LNS11] LIU Z., NAVATHE S. B., STASKO J. T.: Network-based visual analysis of tabular data. *IEEE VAST 3*, c (Oct. 2011), 41–50. 5

[LPB*06] LEE B., PARR C., BEDERSON B., VEKSLER V., GRAY W., KOTFILA C.: Treeplus: interactive exploration of networks with enhanced tree layouts. *IEEE TVCG 12*, 6 (2006), 1414–26. 3

[LSR*09] LEE B., SMITH G., ROBERTSON G., CZERWINSKI M., TAN D. S.: Facetlens: Exposing trends and relationships to support sensemaking within faceted datasets. In *CHI* (2009), pp. 1293–1302. 3

[Mar04] MARCHIONINI G.: From information retrieval to information interaction. In *Advances in information retrieval*, McDonald S., Tait J., (Eds.). Springer Berlin Heidelberg, 2004, ch. From Infom, pp. 1–11. 3

[Mar06] MARCHIONINI G.: Exploratory search: From finding to understanding. *CACM 49*, 4 (2006), 41–46. 3

[MS88] MARCHIONINI G., SHNEIDERMAN B.: Finding facts vs. browsing knowledge in hypertext systems. *Computer 21*, 1 (Jan. 1988), 70–80. 1, 2, 3

[OJ93] O'DAY V. L., JEFFRIES R.: Orienteering in an Information Landscape: How Information Seekers Get From Here to There. In *Proc. ACM CHI* (1993), pp. 438–445. 1, 2

[RBSW01] RODDEN K., BASALAJ W., SINCLAIR D., WOOD K.: Does Organisation by Similarity Assist Image Browsing? In *Proc. ACM CHI* (2001), pp. 190–197. 3

[TAAK04] TEEVAN J., ALVARADO C., ACKERMAN M. S., KARGER D. R.: The Perfect Search Engine Is Not Enough : A Study of Orienteering Behavior in Directed Search. In *Proc. ACM CHI* (2004), pp. 415–422. 1, 2, 3

[TFP06] TONG H., FALOUTSOS C., PAN J.: Fast Random Walk with Restart and Its Applications. In *Proc. ICDM* (2006), pp. 613–622. 5

[tHPvdW96] TER HOFSTEDE A., PROPER H., VAN DER WEIDE T.: Query formulation as an information retrieval problem. *The Computer Journal 39*, 4 (September 1996), 255–274. 2, 3

[TT73] TULVING E., THOMSON D. M.: Encoding Specificity and Retrieval Processes in Episodic Memory. *Psychological Review 80*, 5 (1973), 352–373. 2, 3

[UJvR06] URBAN J., JOSE J. M., VAN RIJSBERGEN C.: An Adaptive Approach Towards Content-Based Image Retrieval. *Multimedia Tools and Applications 31*, 1 (2006), 1–28. 4

[vHP09] VAN HAM F., PERER A.: "Search, Show Context, Expand on Demand": Supporting Large Graph Exploration with Degree-of-Interest. *IEEE TVCG 15*, 6 (2009), 953–960. 3, 6

[WKDS06] WHITE R., KULES B., DRUCKER S. M., SCHRAEFEL M.: Supporting Exploratory Search. *CACM 49*, 4 (2006), 36–39.

[YSLH03] YEE K.-P., SWEARINGEN K., LI K., HEARST M. A.: Faceted metadata for image search and browsing. In *Proc. ACM CHI* (New York, New York, USA, 2003), ACM Press, p. 401. 3

[ZCCB13] ZHAO J., COLLINS C., CHEVALIER F., BALAKRISHNAN R.: Interactive exploration of implicit and explicit relations in faceted datasets. *IEEE TVCG 19*, 12 (Dec. 2013), 2080–9. 3