Designing Breadth-Oriented Data Exploration for Mitigating Cognitive Biases

Po-Ming Law[*]
Georgia Institute of Technology

Rahul C. Basole[†]
Georgia Institute of Technology

Abstract

Exploratory data analysis involves making a series of complex decisions: what should I explore? what questions should I ask? As users do not have good knowledge about the data they are exploring, making these decisions is non-trivial. In making these decisions, heuristics are often applied, potentially causing a biased exploration path. While breadth-oriented data exploration presents a promising solution to rectifying a biased exploration path, how to design breadth-oriented systems is yet to be explored. In this paper, we propose three considerations in designing systems which support breadth-oriented data exploration. To demonstrate the utility of these design considerations, we illustrate a hypothetical breadth-oriented system. We argue that these design considerations pave the way for understanding how breadth-oriented exploration mitigates biases in exploratory data analysis.

Index Terms: H.5.2 [Information Interfaces and Presentation]: User Interfaces—Graphical user interfaces

1 Introduction

Exploratory data analysis empowers users to discover the unanticipated from data. In search of insights, users examine a body of information and articulate questions about the data in an iterative fashion [8]. Aside from being loaded with a vast amount of new information, users have to make a series of complex decisions while navigating through data: what questions should I ask? which piece of information should I examine to answer my questions? As users do not have good knowledge about the data they are exploring, making these decisions is difficult. Unconscious shortcuts are often applied in making these decisions, letting heuristics to drive users' exploration. While heuristics maintains analysis flow by shielding users from making conscious effort in every step of data exploration, a biased exploration path might hinder insight generation and lead to confirming hypotheses erroneously.

A characteristic in a biased exploration path is lack of breadth. Users may be fixated on a question in the early stage of exploration. As a result, the coverage of a dataset is constrained. For instance, psychology studies show that people have a tendency to search for information which confirms pre-existing hypotheses (confirmation bias) [4]. People also tend to associate higher importance to things they can recall better and potentially explore the related information more (availability heuristics) [9]. Furthermore, data analysis is often initiated by formulating a goal or an anchor. Once the anchor is set, people tend to lean towards it (anchoring bias) [10].

We argue that these bias and heuristics can be alleviated by breadth-oriented data exploration. Take for example anchoring bias. When a user is looking for a car to purchase, he might start by searching for cars with a low price. Without considering other car attributes, he tends to pay excessive attention on the cheap vehicles and makes a suboptimal purchase decision. A breadth-oriented exploration system like Voyager [13] can expose users to the other car attributes, assist users with assessing alternatives which are not as cheap but have other desirable properties and hence help users adjust from the bias.

While systems focusing on breadth-oriented exploration start to emerge, how to design such systems is yet to be explored. Voyager [13] provides insights into designing breadth-oriented systems for exploring tabular data. Yet, a large variety of data types are involved in exploratory data analysis in different domains. Developing new breadth-oriented systems for different data types would be challenging if the design process is not informed by any guidelines.

In this paper, we contribute three considerations involved in designing systems which support breadth-oriented data exploration. To demonstrate the utility of these design considerations, we illustrate a hypothetical system which facilitates breadth-oriented exploration of dynamic networks. Finally, we discuss the challenge, the opportunities and the future work in advancing the science of breadth-oriented exploration.

2 The Information Space Model of Breadth-Oriented Exploration

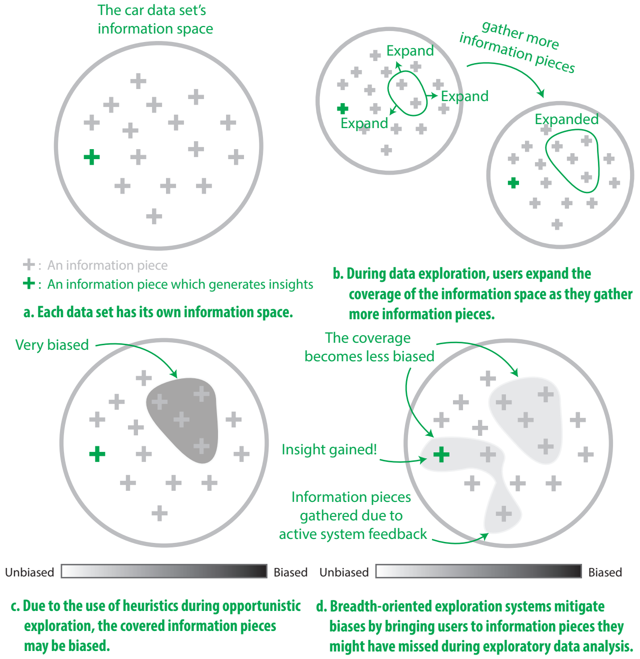
To elucidate the process of breadth-oriented exploration and facilitate the discussion about our three design considerations, we present the information space model.

In the information space model, each dataset has its own information space (Fig. 1a). An information space is a set of information pieces which can be derived from the data. For example, in the notoriously famous car dataset [1], an information piece can be "the dataset covers cars produced between 1970 and 1982" or "acceleration seems to be normally distributed". Some of these information pieces are deemed as insights by a user while some of them are not. Insights are the information pieces which present meaningful knowledge to users, help make decisions (e.g. which car to buy) and validate hypotheses. As insights are user-defined, which information pieces correspond to insights vary among users.

During exploratory data analysis, users expand their coverage of the information space as they gather more information pieces (Fig. 1b). Due to the heuristics they unconsciously resort to during opportunistic exploration, the covered set of information pieces might be biased (Fig. 1c). The consequence is that we might not be able to reach some of information pieces which generate insights. For example, users may be fixated on the price of cars and do not pay attention to the options which are not as cheap but have other good qualities. This can also happen when users constrain themselves to a small set of information pieces in an attempt to confirm their hypotheses and do not explore the alternative hypotheses. Through *active system feedback*, a system which supports breadth-oriented exploration grants users access to the information pieces that they will miss if they explore the information space alone. By bringing users to the unexplored information pieces, breadth-oriented systems help users glean insights they might have missed and mitigate biases of their exploration (Fig. 1d).

[*] e-mail: pmlaw@gatech.edu
[†] e-mail: basole@gatech.edu

[1] https://archive.ics.uci.edu/ml/datasets/auto+mpg


The car data set's information space
+ : An information piece
+ : An information piece which generates insights
Expand
Expand
Expand
gather more information pieces
Expanded
Very biased
The coverage becomes less biased
Insight gained!
Information pieces gathered due to active system feedback
Unbiased
Biased
Unbiased
Biased

a. Each data set has its own information space.

b. During data exploration, users expand the coverage of the information space as they gather more information pieces.

c. Due to the use of heuristics during opportunistic exploration, the covered information pieces may be biased.

d. Breadth-oriented exploration systems mitigate biases by bringing users to information pieces they might have missed during exploratory data analysis.

Figure 1: Four characteristics of the information space model.

3 Three Considerations for Designing Breadth-Oriented Data Exploration

In this section, we offer three considerations involved in designing systems which support breadth-oriented exploration. These design considerations are unit of exploration, user-driven vs system-driven exploration and related vs systematic exploration.

3.1 Unit of Exploration

A unit of exploration is a tool for thinking about the breadth of users' exploration. It can be a dimension or a data case. The use of this tool is twofold. It facilitates designing systems which support exploration in breadth and it serves as a metric for quantifying the effectiveness of a system in promoting breadth-oriented exploration.

If the unit of exploration is dimension, designers may consider providing dimension coverage information to help users keep track of how much of the information space they have explored so far. A dataset can have hundreds of dimensions. While analyzing a dataset, users can explore any combinations of these dimensions. If users do not know what combinations have or have not been seen, the breadth of their exploration can be limited. Showing dimension coverage information creates the awareness of what has not been explored to steer users towards the unexplored dimensions and combinations. Sarvghad et al. [7] demonstrated that incorporating dimension coverage information into an interface increases breadth of exploration without sacrificing depth. Rather than showing information about users' provenance of exploration, a breadth-oriented system may actively present more dimensions to users while they are exploring the data. For instance, as users indicate their interests in one dimension, Voyager [13] expands the covered information pieces in the information space by displaying statistical charts with unseen dimensions. Theoretically, similar principles can be applied when a data case is a unit of exploration. For example, a breadth-oriented system can be designed to help users track what data cases (e.g. documents) have already been explored and what have not. It is similar to email applications which allow us to mark an email as "read". We are more aware of the unread emails consequently.

Apart from helping designers think about how an interface should be designed to encourage breadth-oriented exploration, unit of exploration also provides a simple metric of breadth of exploration. When designers evaluate the effectiveness of their systems in encouraging broad exploration, they can measure how many dimensions or data cases the subjects covered while using their systems in comparison with other tools which do not support breadth-oriented exploration. The assumption is that the more the units of exploration covered, the greater the breadth of exploration. While this metric is simple, it might not be as reliable as other metrics such as the number of findings which indicate a new line of inquiry during exploratory analysis [7].

3.2 User-driven vs System-driven Exploration

Another question in designing breadth-oriented exploration system is whether the expansion of information space is user-driven or system-driven.

User-driven systems create the awareness that users' exploration might be biased. Having known that their exploration has been biased, users can expand the information space in a less biased way. Albeit focusing on cohort selection rather than data exploration, adaptive contextualization [2] illustrates how creating awareness of bias can help users adjust from bias. As users select a patient cohort, the system presents information about the distribution of the selected cohort. Being aware of the bias in the distribution, users can adjust the criteria for cohort selection. In the context of data exploration, system designers can create the awareness of a biased exploration path by presenting information about what has been explored so far, how much has been explored so far and even what other people have explored (like scented widget [12]).

Different from user-driven breadth-oriented exploration, in system-driven exploration, systems actively expand the information space while users are exploring the data. As users express their interests in something (e.g. a dimension), these systems actively present something related but different. With Voyager [13], users can express their interest in some dimensions and the system displays charts with the selected dimensions as well as an unseen dimension. This technique is widely adopted by the graph visualization community (e.g. [1, 3, 11]). For example, with Apolo [1], users start the exploration by putting some nodes into groups. The system then searches for some nodes which are related to the group from a network with thousands of nodes and present them to users.

The key difference between user-driven exploration and system-driven exploration lies in what information is presented to users. In user-driven exploration, systems present information about users' exploration history to create an awareness that users' exploration may be biased. Users rather than systems are responsible to adjust their exploration. In system-driven exploration, systems present extra information extracted from the dataset to directly expand the coverage of the information space. Adjusting users' exploration is the responsibility of systems rather than users.

3.3 Related vs Systematic Exploration

If the expansion of information space is driven by systems, two more considerations are involved: whether the expansion is based on users' interests (related exploration) or the expansion is systematic and not related to users' interests (systematic exploration).

Both Voyager [13] and Apolo [1] mentioned in the previous paragraph falls into the category of related exploration. There are two sides of this same coin: presenting something *similar* to but *different* from users' interests. By showing something *similar* (like Voyager [13] which shows statistical charts related to the variables users are interested in), these systems maintain users' theme and flow of analysis. By showing something *different* from what is indicated by users (like Voyager [13] which shows some unseen dimensions in the recommended statistical charts), these systems expand the information space, bring users to information pieces

Unit of Exploration	User-driven vs System-driven Exploration	Related vs Systematic Exploration
Types: dimension or data case **Uses:** **1)** facilitating the design of breadth-oriented exploration systems (e.g. when the unit of exploration is dimension, systems can provide dimension coverage information to create awareness of biases). **2)** quantifying the effectiveness of a system in promoting breadth-oriented exploration.	**User-driven:** systems create awareness of a biased exploration path and users expand the information space in a less biased manner as they become aware of their biases. **System-driven:** systems actively present information extracted from the dataset to expand the information space in a less biased manner.	**For system-driven exploration:** **Related:** As users express their interests, systems present something related but different. **Systematic:** To complete an analysis, users need to finish some steps in a predefined exploration path.

Table 1: The three considerations for designing breadth-oriented data exploration.

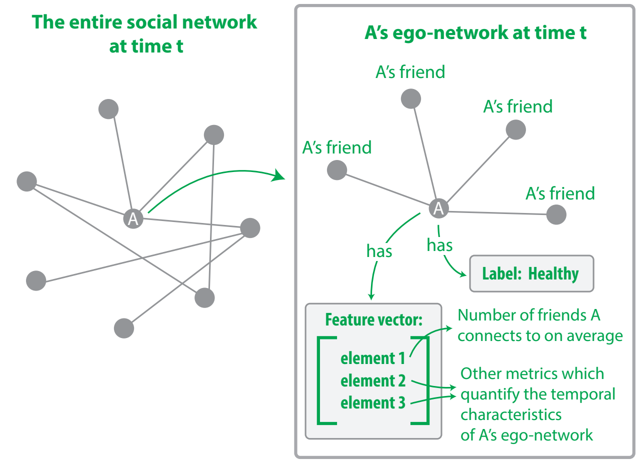

The entire social network at time t
A
A's ego-network at time t
A's friend
A's friend
A's friend
A's friend
A
has
has
Label: Healthy
Feature vector:
element 1
element 2
element 3
Number of friends A connects to on average
Other metrics which quantify the temporal characteristics of A's ego-network

Figure 2: Mary's dynamic network dataset. Note that both the entire network (left) and A's ego-network (right) change over time.

which he might have missed and mitigate biases in users' paths of exploration. This approach bears some resemblance with Amazon which recommends products that users might have missed based on their purchase history.

For systematic exploration, systems expand the information space in a systematic way. A predefined path of exploration is planned out prior to analysis. In order to complete an analysis, users need to finish all the steps in the predefined path. For example, Perer and Shneiderman [6] concluded that there are 7 steps in social network analysis from their experience with domain experts. Although it might lack flexibility of freely exploring the data, this approach ensures that users will not miss any important information pieces. As users' exploration is driven by predefined paths rather than heuristics, it is likely an effective approach to mitigating biases in data exploration. Yet, designing the predefined exploration path is clearly not an easy task.

4 Application of the Three Design Considerations

To demonstrate the utility of the three design considerations, we propose a hypothetical system which facilitates breadth-oriented exploration of dynamic social networks. In conducting the design study, we first consider a common task involved in social network analysis (SNA). We then illustrate a motivating usage scenario of the proposed system. Finally, we explain how the system is designed based on our three considerations.

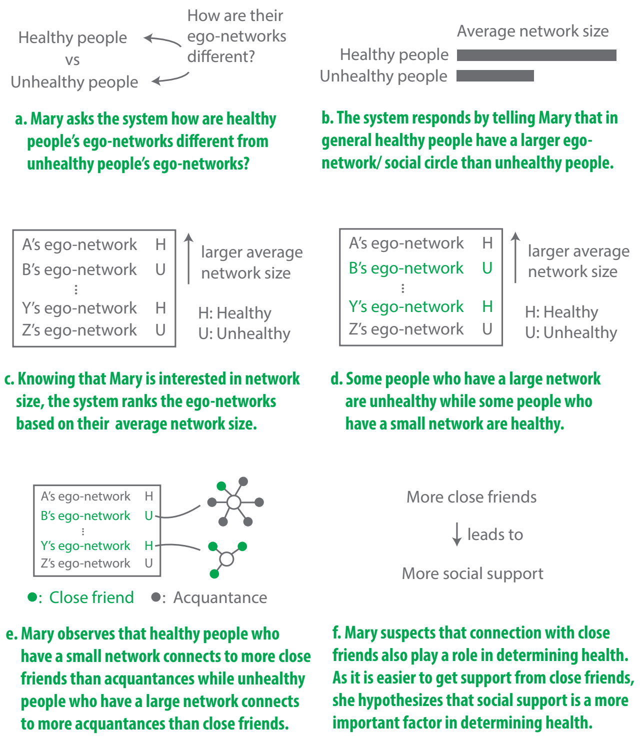
4.1 Task Analysis

One common task in social network analysis is to understand the temporal characteristics of different groups of ego-networks. In an online social network, each person in this network is connected to many other people (i.e. their friends). An ego-network consists of a focal node (a person) and the nodes which are directly connected to it (the person's friends). These ego-networks are dynamic in nature because the set of nodes which are connected to a focal node changes over time.

Consider Mary, a healthcare researcher who wants to explore a dynamic social network. In healthcare domain, researchers like Mary are interested in knowing whether a larger social network leads to better health [5]. In Mary's dataset (Fig. 2), each node is described by a label (either healthy or unhealthy) and a feature vector which quantifies the temporal characteristics of the node's ego-network. Suppose there are three elements in this feature vector: average size of the node's ego-network (average number of friends connected to the focal node), a metric which indicates how fluctuating the ego-network size is and a metric which indicates the average number of clusters in the ego-network. Our goal is to design a system which enables Mary to make sense of the temporal characteristics of the ego-networks of healthy and unhealthy people. This system should encourage Mary to explore her dataset in breadth and mitigate her biases during data exploration.

4.2 Usage Scenario

Mary recalls that many of her friends who look healthy have a large social circle. She has a gut feeling that a larger social circle leads to better health (availability heuristics). She initiated her analysis with the system by searching for evidence to confirm her hypothesis (confirmation bias). To do so, she asks the system what are the distinguishing features of the healthy group compared with the unhealthy group (Fig. 3a). Using data mining techniques, the system tells her that the healthy group in general has a larger social network while the unhealthy group in general has a smaller social network (Fig. 3b). A system which does not support breadth-oriented exploration will stop the analysis here, causing Mary to believe that her hypothesis is true. Knowing that Mary are interested in ego-networks with a large average network size, the proposed breath-oriented exploration system ranks the ego-networks based on their average network size (i.e. the average number of nodes to which a focal node is connected) (Fig. 3c). Mary notices that some people who have a large average network size are unhealthy and some people who have a small average network size are healthy (Fig. 3d). She visualizes these ego-networks using a node-link diagram. She observed that the unhealthy people who have a large network in general connect to more acquaintances than close friends and the healthy people who have a small network in general connect to more close friends than



a. Mary asks the system how are healthy people's ego-networks different from unhealthy people's ego-networks?

b. The system responds by telling Mary that in general healthy people have a larger ego-network/ social circle than unhealthy people.

c. Knowing that Mary is interested in network size, the system ranks the ego-networks based on their average network size.

d. Some people who have a large network are unhealthy while some people who have a small network are healthy.

e. Mary observes that healthy people who have a small network connects to more close friends than acquantances while unhealthy people who have a large network connects to more acquantances than close friends.

f. Mary suspects that connection with close friends also play a role in determining health. As it is easier to get support from close friends, she hypothesizes that social support is a more important factor in determining health.

Figure 3: A usage scenario of the proposed breadth-oriented exploration system.

acquaintances (Fig. 3e). As it is easier to get social support from close friends than acquaintances, Mary has an alternative hypothesis: social support rather than social network size is a more important factor in determining health (Fig. 3f). The breadth-oriented exploration system successfully helps Mary move beyond the original line of inquiry to consider an alternative hypothesis.

4.3 Designing Based on the Three Considerations

The three design considerations are applied when we design the proposed system.

Unit of exploration: In designing the system for exploring Mary's dynamic network, we can choose node, link or dynamic ego-network as a unit of exploration. To capture the temporal characteristics better, we choose *dynamic ego-networks* as a unit. When Mary expresses her interests in the ego-networks with a large average size, the system presents a ranked list of *dynamic ego-networks* based on their average network size.

System-driven vs user-driven exploration: Rather than creating an awareness of a biased exploration path and letting Mary refine her exploration on her own, the proposed system achieves breadth-oriented exploration by actively presenting an ordered list of ego-networks that Mary may be interested in. As the system actively expands the information space by showing information pieces that users might have missed, the system is designed to be *system-driven*.

Related vs systematic exploration: The system creates a ranked list of related ego-networks when Mary demonstrates her interests in ego-networks with a large network size. *Related exploration* is adopted when designing the system.

5 Discussion

We end this paper by discussing the challenge, opportunities and future work in advancing the science of breadth-oriented exploration.

5.1 Challenge: Information Overload

An obvious concern with breadth-oriented exploration is information overload. The extra information presented by breadth-oriented systems in users' course of exploration requires extra cognitive effort to process. Worse still, these systems may present irrelevant information to users. The question concerned is how to ensure the relevance of the information to be presented, particularly in system-driven exploration in which systems actively present information pieces from the dataset. A potential solution is to create models of users based on their exploration history. The model may contain information about a user's interests and what she has explored so far. This approach is similar to recommender systems, which infer what users are interested in based on their profiles, search history or purchase history. With users' models, breadth-oriented systems can predict what information users are interested in, prune the search space of information to be presented and provide more relevant information.

5.2 Opportunities: From Cognitive Bias to Big Data

Beyond cognitive bias, breadth-oriented exploration presents excellent opportunities in the era of big data. Consider a data table with millions of rows. It is likely that users will miss a lot of valuable insights during exploratory data analysis. Due to information overload and the complex decisions to be made during exploratory analysis (e.g. how should I proceed, what should I explore next and what questions should I ask), bias might lurk around the analysis. Systems have full knowledge of the data it contains and can perform unbiased computation on the data. A breadth-oriented exploration system can present useful information that users might have missed and at the same time mitigate cognitive biases.

5.3 Future Work: Fitting the Design Considerations into the Research Agenda

Admittedly, the science of breadth-oriented exploration is still in its infancy. Several questions have to be answered before breadth-oriented exploration systems are widely adopted to mitigate biases during data exploration. For instance, what heuristics are resorted to when users explore data? While many heuristics are well-studied in psychology, how users might apply them while navigating through data is less explored. Furthermore, what is the mechanism by which breadth-oriented exploration alleviates the biases caused by the use of these heuristics? What are the best strategies of mitigating biases by utilizing breadth-oriented exploration? We do not have answers to these questions but we believe that developing and evaluating more breadth-oriented systems is crucial in answering these questions. Our design considerations can provide a starting point for system designers to explore the design space of breadth-oriented data exploration.

References

[1] D. H. Chau, A. Kittur, J. I. Hong, and C. Faloutsos. Apolo: making sense of large network data by combining rich user interaction and machine learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 167–176. ACM, 2011.

[2] D. Gotz, S. Sun, and N. Cao. Adaptive contextualization: Combating bias during high-dimensional visualization and data selection. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*, pp. 85–95. ACM, 2016.

[3] S. Kairam, N. H. Riche, S. Drucker, R. Fernandez, and J. Heer. Refinery: Visual exploration of large, heterogeneous networks through associative browsing. In *Computer Graphics Forum*, vol. 34, pp. 301–310. Wiley Online Library, 2015.

[4] R. S. Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175, 1998.

[5] A. J. Omalley, S. Arbesman, D. M. Steiger, J. H. Fowler, and N. A. Christakis. Egocentric social network structure, health, and prosocial behaviors in a national panel study of americans. *PLoS One*, 7(5):e36250, 2012.

[6] A. Perer and B. Shneiderman. Systematic yet flexible discovery: guiding domain experts through exploratory data analysis. In *Proceedings of the 13th international conference on Intelligent user interfaces*, pp. 109–118. ACM, 2008.

[7] A. Sarvghad, M. Tory, and N. Mahyar. Visualizing dimension coverage to support exploratory analysis. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):21–30, 2017.

[8] E. G. Toms. Information interaction: Providing a framework for information architecture. *Journal of the Association for Information Science and Technology*, 53(10):855–862, 2002.

[9] A. Tversky and D. Kahneman. Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, 5(2):207–232, 1973.

[10] A. Tversky and D. Kahneman. Judgment under uncertainty: Heuristics and biases. In *Utility, probability, and human decision making*, pp. 141–162. Springer, 1975.

[11] F. Van Ham and A. Perer. search, show context, expand on demand: supporting large graph exploration with degree-of-interest. *IEEE Transactions on Visualization and Computer Graphics*, 15(6), 2009.

[12] W. Willett, J. Heer, and M. Agrawala. Scented widgets: Improving navigation cues with embedded visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1129–1136, 2007.

[13] K. Wongsuphasawat, D. Moritz, A. Anand, J. Mackinlay, B. Howe, and J. Heer. Voyager: Exploratory analysis via faceted browsing of visualization recommendations. *IEEE transactions on visualization and computer graphics*, 22(1):649–658, 2016.