# The Role of Interactive Biclusters in Sensemaking

**Maoyuan Sun**
Virginia Tech
Blacksburg, VA USA
smaoyuan@vt.edu

**Lauren Bradel**
Virginia Tech
Blacksburg, VA USA
lbradel1@vt.edu

**Chris North**
Virginia Tech
Blacksburg, VA USA
north@vt.edu

**Naren Ramakrishnan**
Virginia Tech
Blacksburg, VA USA
naren@vt.edu

## ABSTRACT
Visual exploration of relationships within large, textual datasets is an important aid for human sensemaking. By understanding computed, structural relationships between entities of different types (e.g., people and locations), users can leverage domain expertise and intuition to determine the importance and relevance of these relationships for tasks, such as intelligence analysis. Biclusters are a potentially desirable method to facilitate this, because they reveal coordinated relationships that can represent meaningful relations. Bixplorer, a visual analytics prototype, supports interactive exploration of textual datasets in a spatial workspace with biclusters. In this paper, we present results of a study that analyzes how users interact with biclusters to solve an intelligence analysis problem using Bixplorer. We found that biclusters played four principal roles in the analytical process: an effective starting point for analysis, a revealer of two levels of connections, an indicator of potentially important entities, and a useful label for clusters of organized information.

### Author Keywords
Biclustering; Visual Interaction; Intelligence Analysis.

### ACM Classification Keywords
H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

## INTRODUCTION
Intelligence analysts face complex and difficult challenges in analyzing large, unstructured, text datasets. They often seek the help of visual analytic technology, combining the computational benefits of statistical models and knowledge discovery algorithms with the cognitive abilities of humans, to support sensemaking [7]. The challenge is to design a system that integrates these areas, creating efficient visualizations to present computed, structural relationships, and a usable workspace for analysts to perform sensemaking.

Identifying coordinated relationships from text (e.g., find three people who have all visited the same four cities on the same set of days) is a common problem in intelligence analysis. We posit that integrating biclusters into a spatial sensemaking workspace provides a potential solution to identify these connections. A bicluster is a complete bipartite graph where every vertex of one set is connected to all vertices of another set [6], which can be viewed as a bundling of individual relationships into a pair of sets. There are three steps to compute biclusters from the text dataset. First, extract useful entities (e.g., people, location, date) from the text, and then establish relationships between individual entities based on co-occurrence (e.g., a person at a location). Finally, biclusters are formed by grouping entities that share the same relationships (e.g., three people at the same four locations) with algorithms such as CHARM [10]. Spatial workspaces provide a flexible and expressive visual medium to support sensemaking [1]. For example, analysts can organize documents in spatial structures to help synthesize hypotheses. Integrating biclusters into the spatial workspace can enable users to interact with them in a facile way within their analytic process.


*CHI'14*, April 26–May 1, 2014, Toronto, Canada.

http://dx.doi.org/10.1145/2556288.2557337

### Bixplorer
Bixplorer is a visual analytics prototype that combines biclustering algorithms with visual representations (linked matrices) within the spatial workspace [3]. Two considerations of the sensemaking workspace motivate the design of Bixplorer. First, the system must provide a sufficiently rich abstraction of information (e.g., biclusters) that can serve as the basis of more detailed navigation and perusal. Second, sensemaking by users' manual exploration must be seamlessly integrated with the results of knowledge discovery algorithms.

Bixplorer has three major views: a data browser, a preview panel and a spatial workspace, shown in Figure 1. The data browser allows users to search and browse lists of documents, biclusters and entities. The preview panel previews the browsable content, and the workspace enables the visual organization of information. In total, 18 interactions in 6 categories are provided in Bixplorer: 1) ***Open*** documents and biclusters of interest in the data browser and move them to the workspace. 2) ***Remove*** unwanted documents and biclusters from the workspace. 3) ***Show*** related biclusters or documents from a selected document or bicluster. 4) Create user defined ***Links*** between biclusters and documents. 5) ***Extract*** a row or a column of interest from a bicluster. 6) ***Save*** the workspace.

The workspace shown in Figure 1 presents a typical usage scenario in Bixplorer, in which there are 2 biclusters and 5 documents. Bic 40 shows coordinated relationships between 3 dates and 3 people. Based on fbi19, "25 April, 2003"was the date of this FBI report, "23 April, 2003"was the date that person2 bought the train ticket, and "29 April, 2003"was the day that these 3 people planned to meet on the train. Bic 105 shows coordinated relationships between people and organizations, in which AMTRAK was the train that the 3 people

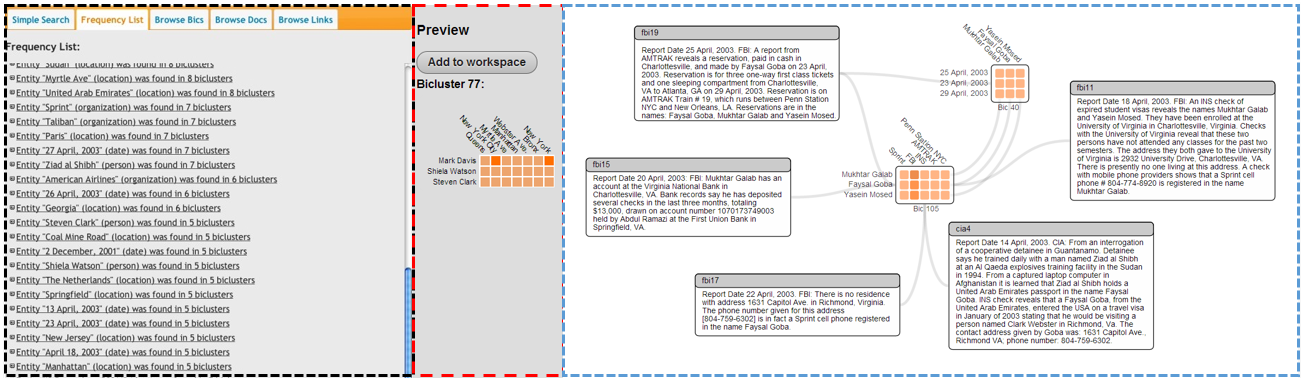


**Figure 1. Three views in Bixplorer: the data browser (left), the preview panel (middle) and the spatial workspace (right).**

would meet, Penn Station NYC was a train station, and all others were organizations in US. Based on the important plots reported in cia4 (person4 received explosive training before, and entered USA with a United Arab passport in the name of person1), and fib11 (person1 and person3 were registered as students in UVA, but they never came to any classes), analysts may make a hypothesis that a planned explosive attack would be performed by the three people on 29 April, 2003 at the AMTRAK train #19.

Previously we discussed some preliminary results of features usage in Bixplorer (e.g., number of used biclusters and user-generated links) [3]. However, the following questions still remain unanswered: How are biclusters integrated into the sensemaking process and used within the sensemaking loop? Do biclusters serve as a helpful guidance for foraging related information or a meaningful summary for synthesized information? To answer these questions and to inform the design of future visual analytics tools, we conducted a user study.

## RELATED WORK

Many visual analytic tools allow users to explore individual relationships, based on text co-occurrence, between entities extracted from raw text documents. For example, Jigsaw's List View [9] visually represents relationships similarly to parallel coordinates, whereas Analyst Notebook and Entity Workspace [2] use entity-relationship network diagrams. Although they provide solutions to visually represent individual relationships, these tools have limited capability for coordinated relationships discovery. Too many edges between entities visually obscure coordinated relationships. Besides, filtering edges by interactively selecting entities requires users to manually search all possible combinations of entities to find the combinations that share common relationships.

Biclusters, incorporated in visual analytics tools, provide a potential solution, which has been explored in the realm of bioinformatics and social network analysis. BicOverlapper [8] uses a bubble map to infer similarities between biclustering results for microarray analysis. BiCluster Viewer [4] applies biclusters in the form of matrix diagrams to assist gene expression data analysis. NodeTrix [5] facilitates exploration of social networks (e.g., co-authorship) through a hybrid visualization that combines adjacency matrices for dense subgraphs, similar to biclusters, with node-link diagrams for sparse connections between subgraphs. Although biclusters appear to be useful for sensemaking tasks, there are few studies on how people use biclusters in human analytic process.

## METHODOLOGY

The user's task in this study was to analyze a set of fictitious intelligence reports, identify any planned attacks, and hypothesize about the terrorist plot. The dataset used in this study (called Sign of the Crescent) contains 41 documents regarding a coordinated terrorist plot in three US cities. 24 reports are relevant to the plot and 17 reports are irrelevant. We used CHARM [10] to generate biclusters from the dataset with the "support"parameter set to 3, which assured that each bicluster has at least 3 rows and 3 columns. This resulted in 284 unique entities, 495 relationships, and 109 biclusters. These settings were selected to match the needs of typical text analysis scenarios.

This study included 15 participants between the age of 20 and 27, and each was compensated for participation. 4 participants were graduate students, and 11 participants were undergraduate students. 6 participants had prior experience with visual analytics tools, but none had prior experience with biclusters. The study procedure consisted of three parts. First we explained the nature of the dataset to users, and used a separate dataset to demonstrate all features in Bixplorer without instructions on any particular analytical approach. Second, users were asked to assume the role of an analyst and investigate the dataset for 1.5 hours to identify any threats of attacks. Finally, we asked users to explain their findings, and interviewed them on their use of Bixplorer to find how biclusters impacted their analysis. Data were collected from log files, workspace screenshots, observations and interviews. Three components of each interaction were logged: the time stamp, interaction type and the target object (e.g., a bicluster or a document that is interacted with by users).

## RESULTS

We found that biclusters in Bixplorer played four principal roles in intelligence analysis: 1) an effective starting point for analysis, 2) directing attention toward connections at both the micro and macro level, 3) an indicator of potentially important entities, and 4) a useful label for clusters of organized

| User | Number of docs opened by users before opening their 1st bic | Timestamp users opened the 1st bic (min:sec) | Relevance of the 1st doc opened before / after users opening their 1st bic | Percent of bic based interactions used to find relevant docs | Percent of bics with non-uniform color in the workspace | Identified attack number |
|---|---|---|---|---|---|---|
| U1 | 6 | 41 : 52 | irrelevant / relevant | 51% | 68% | 1 |
| U2 | 1 | 52 : 31 | irrelevant / irrelevant | 34% | 73% | 1 |
| U3 | 0 | 01 : 19 | none / relevant | 79% | 78% | 2 |
| U4 | 3 | 04 : 11 | relevant / relevant | 55% | 76% | 1 |
| U5 | NA | NA | NA | NA | 100% | 2 |
| U6 | 0 | 01 : 09 | none / relevant | 64% | NA | 1 |
| U7 | 0 | 04 : 10 | none / relevant | 100% | 93% | 2 |
| U8 | 0 | 01 : 08 | none / relevant | 89% | NA | 2 |
| U9 | NA | NA | NA | NA | 80% | 1 |
| U10 | 1 | 01 : 22 | relevant / irrelevant | 86% | 85% | 1 |
| U11 | 0 | 00 : 52 | none / relevant | 78% | 62% | 2 |
| U12 | 0 | 01 : 15 | none / relevant | 75% | 92% | 1 |
| U13 | 0 | 00 : 50 | none / relevant | 35% | 82% | 1 |
| U14 | 6 | 29 : 31 | irrelevant / relevant | 10% | 0% | 1 |
| U15 | 0 | 02 : 45 | none / relevant | 48% | 100% | 2 |

**Table 1. Summary of user data in the study. Denote: bic(s) for bicluster(s) and doc(s) for document(s).**

information. Table 1 summarizes the data in this study. Unfortunately, log files for users 5 and 9 (henceforth, U5 and U9) and workspace screenshots for users 6 and 8 (henceforth, U6 and U8) were corrupted, so we excluded U5 and U9 for log analysis, and excluded U6 and U8 for workspace analysis. Overall, using Bixplorer, all users were able to identify at least one of the 3 hidden attacks and the coordinator of the 3 attacks, although none found all 3 attacks.

### An Effective Starting Point for Analysis

Biclusters quickly direct users to important relevant documents as a starting point for analysis. In this study, we did not provide any instructions about how to start analysis, so all users started their analysis based on their preference. Based on the second column in Table 1, 5 of 13 users began by reading documents; The other 8 users began by examining biclusters. Based on the third column, 2 (U4 and U10) of the 5 who began with documents quickly abandoned this approach within 5 minutes and switched to using biclusters. This indicates that there was some preference for starting with biclusters.

For 7 of the 8 users (except U2) who began with biclusters, the documents that they opened based on their first bicluster were relevant to the solution. We observed that 5 of these users used their first bicluster to identify important entities that they then used to search for relevant documents. The other 2 used the bicluster to directly show relevant documents. Of the 5 users who did not begin with biclusters, 3 began with irrelevant documents because they began reading documents in an arbitrary order. These 3 users persisted reading documents for about 30 minutes or more before opening biclusters. This trend indicates that biclusters provided an efficient way to find relevant documents at the beginning of analysis.

### Directing Attention Toward Connections

On the micro level, each bicluster revealed a set of relationships between entities of two groups. For example, U13 mentioned that he used biclusters "*to see the relationships between items within*". On the macro level, a bicluster could indicate semantic connections among shared lexical items in multiple documents. Micro level connections led users to explore macro level connections, thus enabling them to connect more detailed information from the documents and synthesize semantic hypotheses. For example, U2 mentioned using biclusters "*to find connections between news [documents] and people, places and news [documents]*". In Bixplorer, users can perform eight different interactions with a bicluster (e.g., extract a row), only one of which is used to show documents related to the bicluster. Based on the fifth column in Table1, most of the users' interactions with biclusters were to find documents and therefore better understand the detailed connections between them. Also, through biclusters, users connected documents together frequently by placing them near each other in the spatial layout and occasionally with user-defined links. Taken together, biclusters helped users perceive connections at both levels: co-occurrence of entities at the micro level, and relevance for documents at the macro level.

### An Indicator of Potentially Important Entities

Biclusters enhance the capability of users to recognize potentially important entities. Each bicluster in Bixplorer is represented as a grid of cells, and each cell presents a relationship between two entities. The cell color indicates the frequency of co-occurrence for the two related entities. Thus, uniform coloring of cells within a bicluster means that all relationships within this bicluster have the same frequency, and non-uniform coloring means that at least one relationship has different frequency. An example of uniform and non-uniform biclusters in Bixplorer is shown in Figure 2.

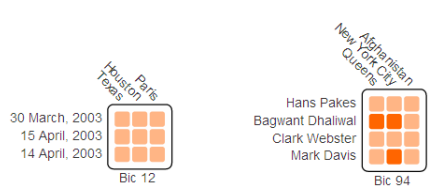


**Figure 2. A uniform bicluster (left) and a non-uniform bicluster (right).**

For the Crescent dataset, the number of non-uniform biclusters found by Bixplorer happens to be 72% of the total number of biclusters found. From the sixth column in Table 1, the non-uniform biclusters rate in 10 users' workspace is higher than 72%. Thus, the ratio of non-uniform to uniform biclusters found on these users' workspaces is greater than the ratio of non-uniform to uniform biclusters found in the dataset. This indicates that users preferred non-uniform over uniform biclusters in their analysis. Some users interpreted that high frequency relationships in non-uniform biclusters indicate important entities. U10 mentioned that "*I mostly focused on biclusters with a large variation in color, since those biclusters usually yielded the strongest correlations*". Although the rate of non-uniform biclusters in U11's workspace is lower than 72%, U11 still admitted that "*The ability to see how many documents connect each two pieces of information just by looking at the shade of the color was a useful thing for quick glances*". U14 removed almost all biclusters (except one) from the workspace at the end of the study, so the rate of non-uniform biclusters in U14's workspace is 0%. In fact, in this dataset, the high frequency cells typically represented the important persons who organized the correlated activity.

### A Useful Label for Clusters of Organized Information

According to the screenshots of their final spatial workspaces, users frequently used biclusters as a label for the surrounding spatially organized information. A typical example of such spatial layout from U13's workspace is shown in Figure 3. A bicluster with one column and one row is placed in the center as a semantic label for the cluster of eight biclusters. This special bicluster was created by U13 using the *extract* interaction. In this bicluster, the row name is "Mark Davis", and the column name is "Queens". All surrounding biclusters list more detailed relevant relationships associated with "Mark Davis"and/or "Queens".

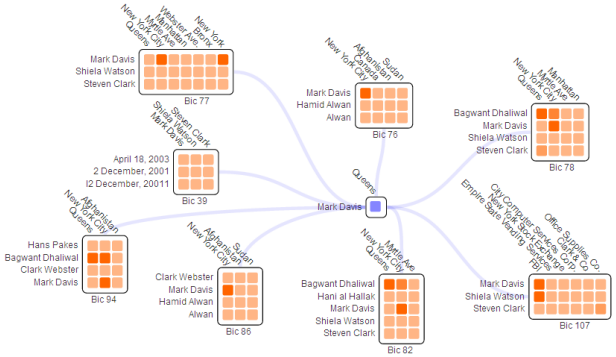


**Figure 3. Parts of spatial layout in U13's workspace. The bicluster in the center presents the relationship between "Mark Davis"and "Queens".**

In fact, 7 of 13 users' final workspaces had spatial layouts similar to that shown in Figure 3, and 6 of these 7 users mentioned that biclusters provided the labeling capability to assist their analysis. For example, U3 mentioned that "*[I] chose one [bicluster] that had. . . , something that sounded terrorist-like*"and U7 said that, "*[I] used them [biclusters] to title clusters of documents*". By referring to a bicluster as a label, users can quickly retrieve useful information from the spatial layout. For example, when reporting his findings, U12 looked at a specific bicluster in his workspace and then mentioned that "*[personA] has connections to [personB] who allegedly visited [personC] at a non-existent address. . .*" From U12's workspace screenshot, we found that this bicluster presents a collection of information related to *personA*.

## CONCLUSION

A bicluster in Bixplorer is a highly abstracted set of relationships. Considering the sensemaking loop, biclusters are useful for foraging information in three ways: directing users to relevant documents as a starting point, discovering potential connections of both the micro level and the macro level, and indicating potentially important entities. Biclusters are also helpful for synthesis by labeling organized information.

## ACKNOWLEDGMENTS

This research was partially supported by NSF grant #CCF-0937133.

## REFERENCES

1. Andrews, C., Endert, A., and North, C. Space to think: large high-resolution displays for sensemaking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (2010), 55–64.
2. Bier, E. A., Ishak, E. W., and Chi, E. Entity workspace: an evidence file that aids memory, inference, and reading. In *Intelligence and Security Informatics*. Springer, 2006, 466–472.
3. Fiaux, P., Sun, M., Bradel, L., North, C., Ramakrishnan, N., and Endert, A. Bixplorer: Visual analytics with biclusters. *Computer 46*, 8 (2013), 90–94.
4. Heinrich, J., Seifert, R., Burch, M., and Weiskopf, D. Bicluster viewer: a visualization tool for analyzing gene expression data. In *Advances in Visual Computing*. Springer, 2011, 641–652.
5. Henry, N., Fekete, J.-D., and McGuffin, M. J. Nodetrix: a hybrid visualization of social networks. *Visualization and Computer Graphics, IEEE Transactions on 13*, 6 (2007), 1302–1309.
6. Jin, Y., Murali, T., and Ramakrishnan, N. Compositional mining of multirelational biological datasets. *ACM Transactions on Knowledge Discovery from Data (TKDD) 2*, 1 (2008), 2.
7. Pirolli, P., and Card, S. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of International Conference on Intelligence Analysis*, vol. 5 (2005), 2–4.
8. Santamaría, R., Therón, R., and Quintales, L. Bicoverlapper: a tool for bicluster visualization. *Bioinformatics 24*, 9 (2008), 1212–1213.
9. Stasko, J., Görg, C., and Liu, Z. Jigsaw: supporting investigative analysis through interactive visualization. *Information visualization 7*, 2 (2008), 118–132.
10. Zaki, M. J., and Hsiao, C.-J. Charm: An efficient algorithm for closed itemset mining. In *SDM*, vol. 2 (2002), 457–473.