EXPLORING THE AUTOMATIC RECOGNITION OF PIE CHART INFORMATION MESSAGES

Eric Balawejder, Tyler Traub, Richard Burns
West Chester University
eb621515@wcupa.edu, tt840371@wcupa.edu, rburns@wcupa.edu

ABSTRACT
Information graphics, such as pie charts, bar charts and line graphs, are frequently used to convey high-level intended messages in popular media. In this paper, we focus on how high level intended messages are communicated in pie charts. We have assembled a corpus of pie charts collected from popular media. We then annotated each pie chart in the corpus with its high-level intended message that we recognized for it. In this paper, we present our classification of the kinds of high-level messages that we discovered for pie charts into message categories, and present some of the communicative signals that we observed, which assisted us in recognizing their communicative messages. Finally, we describe a Bayesian network implementation that works towards the end-goal of our research: building a system that automatically hypothesizes the intended message of a pie chart.

KEYWORDS
Information graphics, pie charts, high-level messages, communicative signals, Bayesian Network.

1. Introduction

Multimodal documents often incorporate information graphics, such as bar charts and pie charts, alongside article text to achieve a set of communicative goals [7] [5]. In popular media (magazines such as *Time* and newspapers such as *USA Today,* but not scientific articles), information graphics are sometimes included in an article in order to convey an additional, supplemental high-level message that transcends the low-level data points in the graphic. For example, the grouped bar chart in Figure 1 ostensibly conveys a high-level message that *"Women are more likely than men to delay medical treatment"*, which is more communicative than the lower-level data: that the first bar is 21%, the second bar is 37%, etc.

The idea that information graphics can be considered a form of language follows Clark [3] who noted that language is any "signal" or lack thereof, where a signal is any deliberate action that is intended to convey a message, including gestures and facial expressions. Our view is that information graphics are a form of language, where the designer of a graphic is able to deliberately use communicative signals to help convey an intended message to the viewer of the graphic.

In this paper, we present preliminary results of designing a system that is capable of automatically determining the most likely high-level message of a pie chart by observing the absence or presence of communicative signals in the graphic.

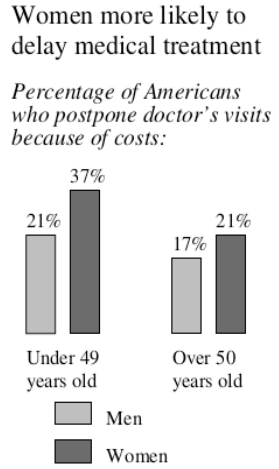

Figure 1: From *USA Today.*

It is non-trivial to identify the intended message of an information graphic. Carberry [2] found that the high-level message of an information graphic is not always written into its caption or accompanying text. This makes it difficult to automatically discover a graphic's high-level message using only natural language processing on the graphic's surrounding text.

This work is the first of our knowledge that studies the problem of recognizing the intended high-level message of a pie chart when it is drawn in popular media.

We have collected a set of pie chart information graphics occurring in popular media, and examined these charts to identify: (1) the types of high-level messages that graphic designers convey using pie charts, and (2) the kinds of communicative signals present in pie charts that appear likely to assist the recognition of high-level messages.

We created a XML representation for pie chart information graphics that fully capture the properties needed (such as number of slices, their sizes, etc.) to reason about the graphic. We then designed and implemented Bayesian network that captures the probabilistic relationship between high-level pie chart messages and their communicative signals. The end-goal of the network is to use posterior communicative evidence to predict the high-level intended message of a new pie chart graphic.

One application of this research is for sight-impaired individuals who cannot view information graphics. Alternative access screen readers can convert the content of a pie chart to text, but only at the level of low-level raw data: "the first pie chart slice is 18.5%, the second pie chart slice is 7.3%, etc." Our research aims to generate the high-level message as text for sight-impaired users. A second application for this work is to use the recognized intended message of a pie chart as an indexing feature in an information retrieval system. One system that can benefit from our work is Zanran (http://www.zanran.com), which takes a user query such as "income" and returns multimodal documents that include information graphics that the system believes are relevant to the query. Currently, the system likely attempts to find relevant information graphics by performing OCR on any text within the graphic; with our system, a retrieval engine could also take into account the high-level message of information graphics.

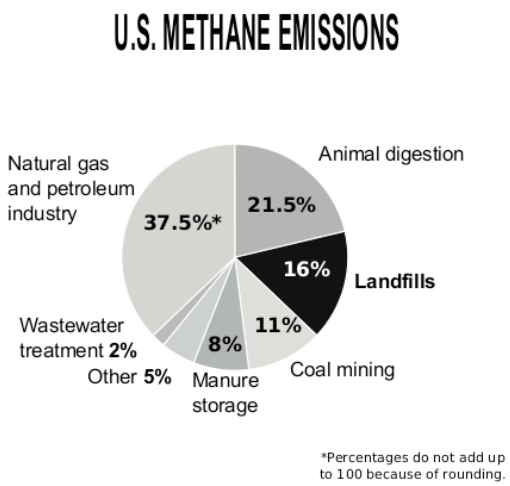

Figure 2: From *National Geographic.*

Section 2 of the paper describes relevant related work. Section 3 describes our collection of pie charts and presents the types of high-level messages that we have recognized in them. Section 4 presents some of the communicative signals that we observed graphic designers use in designing pie charts that have a communicative intent. Section 5 and Section 6 discusses how we represent a pie chart in an XML format so that we can automatically process it. We then explain a Bayesian network framework that we are currently implementing and show how our corpus of pie charts is used to automatically train the conditional probability tables in the Bayesian network. Finally, we discuss the current and future work of this project.

2. Related Work

Previous research has explored automatically recognizing the high-level intended messages of other information graphic types: simple bar charts [4], line graphs [8], and grouped bar charts [1]. These three projects each made use of a Bayesian Network to probabilistically determine the most likely high-level intended message based off identified communicative signals in these graphics. Similar to our project, they first collected a corpus of graphics, which they then used in training their system.

Although our work is similar to these previous studies, each type of information graphic is able to convey a unique set of possible messages. In addition, different communicative signals are utilized by graph designers to help convey the high-level intended messages. Therefore, the end-result for each of the systems has been very different.

3. Pie Chart Message Categories

We collected 115 pie chart information graphs from popular media[1]. Of those, we retained 90 of the charts, as the rest appeared to contain only data, and did not appear to convey any intended message. We then analyzed the corpus to generalize the kinds of high-level intended messages that we recognized in the pie chart graphics into message categories. This section describes and presents examples of some of our identified pie chart message categories.

[1] Our corpus of pie charts is publically available at: http://taz.cs.wcupa.edu/~eb621515/PieCharts

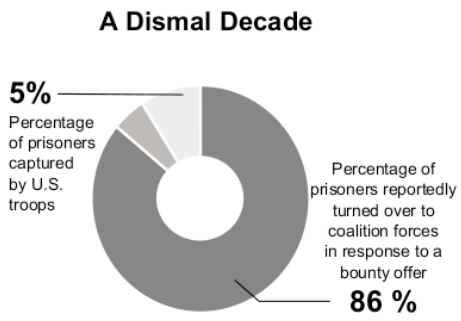

Figure 3: From *Time Magazine.*

There are eleven pie chart message categories that we defined. In this section, we formally define the name of the category, the number of parameters that the category takes, and a short description.

1. *SingleSlice(<s>)*. Single slice messages recognize a high-level message that involves a single, salient, pie chart slice. Generally, the pie charts that fall within this category seem to be designed so that the graph viewer compares a specific, single slice against the other slices in the pie chart. For example, consider the pie chart in Figure 2. This pie chart ostensibly conveys that *"Landfills are a significant source of U.S. methane emissions, the third highest, behind the natural gas and petroleum industry as well as animal digestion"*. The parameter <*s*> in the message category syntax is instantiated with the single pie chart slice that is to be compared against the other slices. That is, this message would be represented as: *singleSlice(s=Landfills)*.

2. *Fraction(<x>)*. The Fraction message describes a visual representation of a fraction of the pie chart. For example, some pie charts that we have collected seem to have a high-level message that some slice is a quarter of the pie, or a half of the pie. The parameter <*x*> is instantiated with a single pie chart slice whose fraction-size of the overall pie is recognized.

3. Versus(<s_1,s_2>). Versus messages capture two salient slices, which are compared against each other. In contrast to single slice messages in which a salient pie chart slice is compared against the rest of the slices in the pie chart, the two salient slices in versus messages are compared with each other rather than the other slices. For example, the pie chart in Figure 3 ostensibly conveys the message that *"most prisoners were turned over to coalition forces because of bounties, rather than being captured by troops"*. The versus message category is instantiated with two parameters: <*s1*> and <*s2*>, the slices that should be compared with each other.

4. BiggestSlice(). Biggest slice messages identify a single slice of the pie chart that is larger than all of the other slices. Because only one slice can be the largest (assuming no ties), the biggest slice message category has no parameters. For example, presumably the intended message in the pie chart in Figure 4 is that *"there were a greater number of male deaths than female deaths in which illicit fentanyl was detected"*.

5. MajoritySlice(). Majority Slice messages represent the recognition that a slice of the pie chart holds additional meaning because it is greater than 50% of the pie chart.

6. AddSlices(< s_1, s_2,….., s_n >). Add Slices messages involve the graph viewer recognizing that the intended message is to aggregate multiple slices of the pie chart together. For example, we have observed a pie chart whose message was to add and recognize the size of three individual slices in the pie chart.

7. TwoTiedForBiggest(< s_1, s_2 >). Two Tied for Biggest messages portray that two pie chart slices are relatively the same size.

8. NoMajority(). No majority messages capture that none of the slices in the pie chart are larger than 50%. Like the biggest slice message category, the no majority message category also has zero parameters. For example, the pie chart in Figure 5 ostensibly intends to convey the high-level message that individuals in search of work take a variable range in time in order to find a job.

9. SmallestSlice(). Smallest Slice messages are the opposite of Biggest Slice messages, where the smallest slice is the most important slice.

Table 1: Types of high-level message categories that we recognized using our collection of pie charts. We have met and discussed consensus annotations for 31 pie charts in our corpus.

Message category	*Description*	*Count*	*Percent*
Single Slice	One slice is more important than the others.	3	9.6%
Fraction	Visual representation of a fraction.	3	9.6%
Versus	Capture two salient slices which are compared against each other.	4	12.9%
Biggest Slice	One slice is bigger than the rest.	3	9.6%
Majority Slice	Slice with most significant meaning that is larger than 50% of the pie chart.	11	35.4%
Add Slices	Add additional slice before comparing.	2	6.4%
Two Tied For Biggest	Compare two slices of equal size < 50%	3	9.6%
No Majority	None of the slices in the pie chart are larger than 50%.	1	3.2%
Smallest Slice	Opposite of Biggest Slice.	0	0%
Close To Half	The most important slice is also within 4% of 50% of the pie chart	0	0%
Number of Parts	The message is simply how many slices are in the pie chart.	1	3.2%

10. CloseToHalf(). Close to Half messages recognize that the biggest slice of a pie is approximately half of the pie chart.

11. NumberOfParts(). This message is simply how many slices are in the pie chart, for example, *"that the pie chart is made up of 4 component pie slices."*

3.1 Annotation and Inter-Coder Agreement

From our categorization of pie chart high-level messages into message categories, we then annotated our corpus using the following procedure: we first individually recognized the intended message for each pie chart and classified it into its appropriate message category. Then, we conducted a consensus-based annotation by meeting as a group and discussing each of our annotations, revising any annotations if we were strongly swayed. The final annotation for each pie chart was decided by majority vote.

So far, we have completed deliberating final annotations for 31 of the pie charts in the corpus, as presented in Table 1. Notably, all of the individual annotators sometimes recognized exactly the same message for a pie chart before any discussion, or a majority of them agreed to exactly the same message after a discussion. This

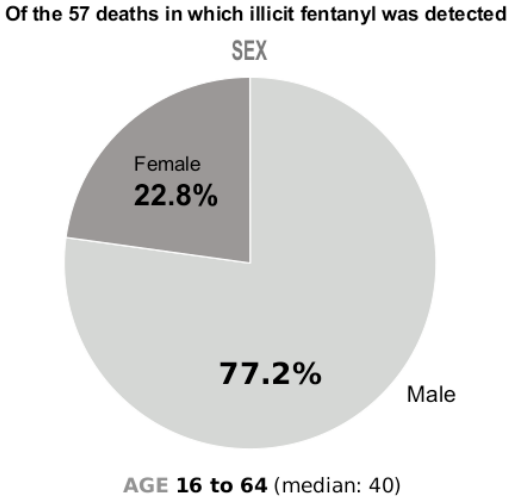

Figure 4: From *The Philadelphia Inquirer.*

level of agreement is a good result and shows that (1) the recognition of pie chart messages is not as subjective as it may initially appear, and (2) our derived and recognized set of pie chart message categories does capture the types of messages that graphic designers convey in popular media using pie charts. A summary of the inter-annotator agreement is in Table 2.

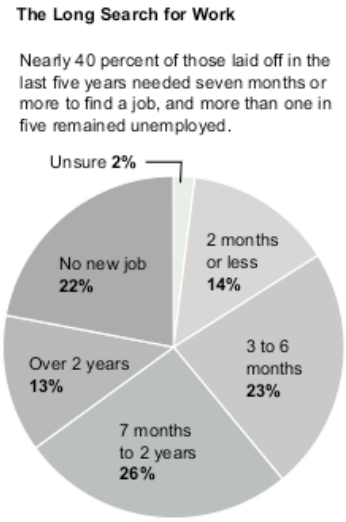

Figure 5: From *The Philadelphia Inquirer*

4. Communicative Signals

We have observed that the presence and absence of communicative signals can assist the recognition of a high-level message that is intended to be conveyed in a pie chart.

4.1 Visual Signals

One visual signal that a graphic designer may use to help communicate some intended message is **prominence**, by coloring a specific pie chart slice a salient coloring, or boldfacing the label of a pie chart slice. An example of this communicative signal is present in Figure 2, which helps signal that *Landfills* should be compared against the other pie chart slices. Another example of a visual signal found in the pie chart corpus is the use of **similar colors** across multiple pie chart slices. For example in Figure 3, the slices for *Bounty* and *Troops* are colored similarly (though not exactly identical), helping signal that they should be compared, while still contrasting them against the Unlabeled 9% slice.[2] Another example of a visual, communicative signal is **separation**, when one pie chart slice is purposely drawn slightly "separated" or "exploded" away from the center of the pie, drawing additional attention to it.

4.2 Linguistic Signals.

Although it does not always fully capture a graphic's intended message, the **caption text** of a pie chart can sometimes serve as a linguistic signal that helps convey its message. For example, in the pie chart in Figure 6, the verb *"split"* helps signal the intended message that there is no majority slice amongst the slices: "will", *"will not"*, and *"unsure"*. We have also observed instances of the article headline of a multimodal article helping to signal the intended message of a pie chart. Another linguistic clue that can serve as a communicative signal is when one pie chart slice is mentioned in the caption or article headline, while the other slices are not mentioned.

[2] In the original graphic, Bounty is colored yellow, Troops is orange, and the unlabeled slice is gray.

Table 2: Summary of the annotation agreement between coders. Table rows display *"The percentage of pie charts that ...*

Percentage	*Description*
36.6%	… all coders recognized with exactly the same message, before any discussion.
56.6%	… a majority of coders recognized with exactly the same message, before any discussion.
63.3%	… all coders recognized with exactly the same message, after discussion.
100%	… a majority of coders recognized with exactly the same message, after discussion.

5. Representing Pie Charts in XML Format

In order perform processing on a pie chart, it needs to first be translated from a potentially noisy graphical format into a representation that fully describes the qualities of a pie chart (its number of slices, their sizes, whether any slice was annotated with a special color to give it

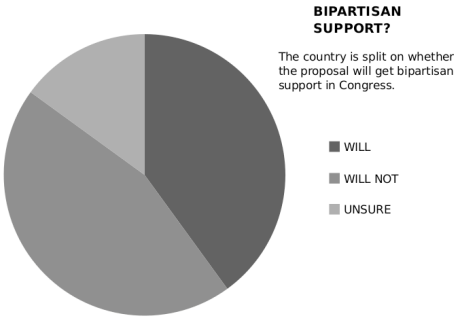

Figure 6: From *USA Today.*

```
<piechart id="P1">
     <!-- Page title: The Afterlife of a
          Landfill
     -->
     <title>US Methane Emissions </title>
        <slice id="1">
                <label>Natural gas and
                    petroleum industry
                </label>
                <annotated_value>37.5%
                </annotated_value>
        </slice>

        ... [omitted to save space]

        <slice id="3">
                <!—
                    slice is highlighted
                     for message content
                -->
                 <highlighted />

                <label>
                   Landfills
                </label>

                <annotated_value>
                   16%
                </annotated_value>
        </slice>
        <slice id="4">
                <label>
                    Coal mining
                </label>
                <annotated_value>
                    11%
                </annotated_value>
        </slice>

        ... [omitted to save space]

        <footnote>
             Percents do not add up
             to 100 because of rounding.
        </footnote>
</piechart>
```

Figure 7: XML Representation of a Pie Chart Graphic

salience, etc.) We chose to represent each pie chart in a custom XML format. Each pie chart in our corpus was hand-translated into a representative XML format. (Interesting research in the area of computer vision, such as the system by Huang et al. [6], can identify an information graphic within a pdf document, perform OCR on it, and then output the sizes of the pie chart slices.) An abbreviated copy of our XML representation for the pie chart in Figure 2 is shown in Figure 7. Note that there are numerous predefined tags to classify and properly describe the data represented in the image. Most of emphasis was placed capturing the qualities of individual slices: their size, any

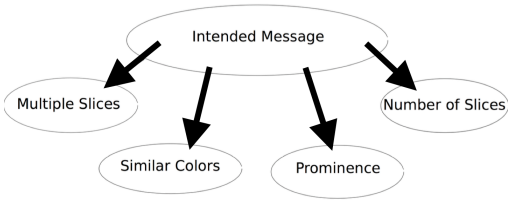

Figure 8: Bayesian Network Design

labeled text, their color, and any highlighting emphasis. We also stored any footnotes that were in the image.

6. Bayesian Network Design

A Bayesian network is a probabilistic graphical model that represents a set of random variables and their conditional dependencies using a directed acyclic graph. It is a model that is frequently used for AI and machine learning applications in business and science. Our current work is the implementation of a Bayesian network that: (1) can be trained on our corpus of pie charts and automatically learn the probabilistic relationships between a pie chart's high-level intended message and its present communicative signals, and (2) can be tested by identifying the communicative signals that are present in a new pie chart, inputting it as evidence in the network, and observing the network's posterior probability of the most likely intended message for that pie chart.

We constructed a Bayesian Network with the following design:

- The intended message of the pie chart is the parent node. This parent node contains the eleven possible intended message categories as states.
- The parent node is directly connected to child nodes, which represent the communicative evidence in the graphic.
- Communicative evidence child nodes include: the number of slices that are in the pie chart (discrete states: 1,2,3,4,5+) whether multiple slices in the pie chart contain more than one descriptive or comparable slice (binary: yes/no), whether there is any pie chart slice that is prominent (binary: yes/no), and whether any two slices have similar colors (binary: yes/no).

This design of the Bayesian network is shown in Figure 8.

Table 3: Learned Conditional Probability Table for the *Number of Slices* Evidence Node

Intended Message	two	three	four	fiveOrMore
singleSlice	42.857	14.286	14.286	28.571
fraction	57.143	14.286	14.286	14.286
versus	25	37.5	25	12.5
biggestSlice	14.286	42.857	28.571	14.286
majoritySlice	40	33.333	13.333	13.333
addSlices	16.667	33.333	33.333	16.667
twoTiedForBiggest	28.571	42.857	14.286	14.286
noMajority	20	20	20	40
smallestSlice	25	25	25	25
closeToHalf	25	25	25	25
numOfParts	20	20	20	40

6.1 Training

We trained the Bayesian network using our corpus of 90 pie charts; the conditional probability tables of the nodes in the network are automatically populated based on the probabilistic relationship between the high-level message categories and the communicative evidence. After training the network, the a priori belief of the system will favor the *Majority Slice* message, as this message category is the most common in our corpus (35.4% as shown in Table 1). The conditional probability tables for each communicative evidence node are also learned. For instance, Table 3 shows the conditional probability table that was automatically learned for the *"number of slices"* node. Observe how the probabilistic relationships for *Single Slice* and *Fraction* seem very different compared to *Biggest Slice.*

6.2 Hypothesizing the Intended Message for a New Pie Chart

After training the network with the 90 instances of pie charts attributes, the goal of the network model is to predict the intended message of a new pie chart graphic that is presented to the network in a csv file format. With the help of the conditional probability structure in the Bayesian Network, this can be accomplished whether the communicative signals are present or not.

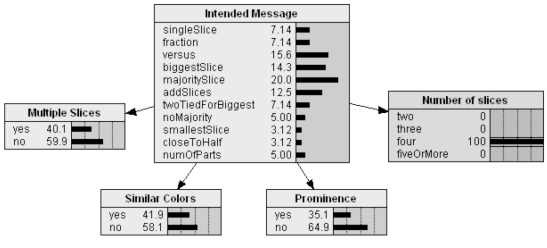

Figure 9: Bayesian Network with Entered Evidence in the *Number of Slices* Node.

When communicative evidence is entered into the child nodes of the network, the root top-level *Intended Message* node is automatically updated with updated probabilities for its belief of the intended message of a new pie chart. For example, Figure 9 demonstrates the entering of *Four Slices* evidence into the child node *Number of Slices*. (Observe that the *Four* state is now 100%, whereas *Two, Three,* and *fiveOrMore,* are 0%, because evidence is now entered.) The probabilities in the *Intended Message* node are then immediately updated. With this entered evidence, the system now believes that the intended message for some pie chart is now *Majority Slice* with a likelihood of 20%. (Note that this likelihood is much lower than the 35.4% likelihood for *Majority Slice* as an a priori probability, as shown in Table 1.)

We are currently working on building additional communicative evidence child nodes into our network, and plan on performing a more thorough evaluation of our system using cross-validation on the graphs in our corpus.

7. Future Work

We have also observed numerous instances of multiple pie charts drawn adjacent to one another, where the single intended message of the graphic seems to involve both pie charts, rather than two individual and separate intended messages. For example, in the multiple pie charts shown in Figure 10, the high-level message conveyed is that the percentage of births to unmarried U.S. women 35 and older increased from 1990 to 2008. This avenue of future work explores the unique types of messages and communicative signals that can be found when *multiple* pie charts are purposely drawn adjacent to each other.

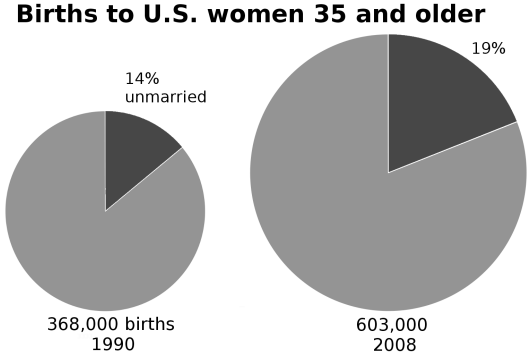

Figure 10: From *National Geographic.*

8. Conclusion

In this paper, we have presented novel research that introduces (1) a corpus of pie charts that we have collected from popular media, (2) a sampling of the types of messages that pie charts are able to convey, (3) examples of communicative signals that help communicate these messages and (4) a Bayesian Network design with the goal of capturing the probabilistically relationship between high-level intended messages and communicative signals. We have shown how evidence entered into the network affects the system's belief of the intended message of a graphic. Compared to other types of information graphics that have been previously studied, the identified messages and communicative signals presented in this paper are unique to pie charts.

References

[1] Burns, R., Carberry, S., Elzer, S., Chester, D.: Automatically recognizing intended messages in grouped bar charts. In: Proc. of the International Conference on the Theory and Application of Diagrams. pp. 8–22 (2012)

[2] Carberry, S., Elzer, S., Demir, S.: Information graphics: An untapped resource of digital libraries. In: Proc. of the Conference on Research and Development on Information Retrieval. pp. 581–588 (2006)

[3] Clark, H.: Using Language. Cambridge University Press (1996)

[4] Elzer, S., Carberry, S., Zukerman, I.: The automated understanding of simple bar charts. Artificial Intelligence 175(2), 526–555 (February 2011)

[5] Green, N.L., Carenini, G., Kerpedjiev, S., Mattis, J., Moore, J.D., Roth, S.F.: Autobrief: an experimental system for the automatic generation of briefings in integrated text and information graphics. International Journal of Human-Computer Studies 61(1), 32–70 (2004)

[6] Huang, W., Tan, C.L.: A system for understanding imaged infographics and its applications. Proceedings of the ACM Symposium on Document Engineering. pp. 9–18. (2007)

[7] Iverson, G., Gergen, M.: Statistics: The Conceptual Approach. Springer-Verlag, New York (1997)

[8] Wu, P., Carberry, S., Elzer, S., Chester, D.: Recognizing the intended message of line graphs. Proceedings of the Conference on Diagrammatic Representation and Inference. pp. 220–234. (2010)