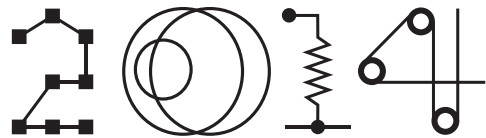


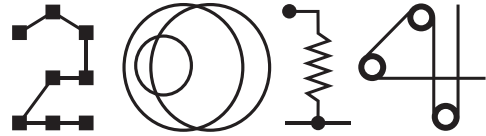
DIAGRAMS
MELBOURNE



Proceedings of the Diagrams 2014 Graduate Symposium

July 28, 2014

DIAGRAMS MELBOURNE



Diagrams are very wide-ranging and open-ended representations that include sketches, drawings, charts, pictures, 2D and 3D geometric models, and maps. Diagrams are a vital tool in human communication in areas such as art and science, as well as commerce and industry. Increased understanding of how effective diagrams can be generated and used has the potential to produce transformative advances in these areas. Research topics include understanding diagrammatic reasoning in humans; understanding the use of diagrammatic representation for communication; developing techniques for automated diagrammatic reasoning; and designing tools for use of diagrammatic representations.

The goal of the Diagrams 2014 Graduate Symposium is twofold. Firstly, the Symposium will provide senior graduate students and recent graduates with the opportunity to present their research and receive feedback from established researchers who will provide comments on each of the presentations. Secondly, the Symposium will provide students with an opportunity to network with each other as future colleagues. The doctoral student symposium will increase the exposure and visibility of young graduate student researchers in these areas, and train them by providing early input and feedback from senior researchers in the field in an interactive and constructive environment.

The Diagrams 2014 Graduate Symposium would like to thank the NSF for its generous support of the symposium and several of its participants under NSF grant no. 1434919.

Graduate Symposium Program

- 1 A Theoretical Approach to Adaptive Visualization
Lydia Byrne, Daniel Angus, Janet Wiles
- 5 Work in Progress: Degree-of-Interest Based Visual Exploration of Heterogeneous Networks
Sanjay Kairam
- 10 Diagrams in Patents: An exploratory introduction
Sylvan G. Rudduck, Mary-Anne Williams, Natalie Stoianoff
- 18 Hunches and Sketches: rapid interactive exploration of large datasets through approximate visualisations
Advait Sarkar, Alan F. Blackwell, Mateja Jamnik, Martin Spott
- 23 Drawing Euler Diagrams and Graphs in Combination
Mithileysh Sathiyarayanan
- 27 Storyline Visualization: Aesthetics and Legibility from Observing Art
Yuzuru Tanahashi and Kwan-Liu Ma

Index of Authors

31

A Theoretical Approach to Adaptive Visualization

Authors: Lydia Byrne¹, Daniel Angus^{1,2}, Janet Wiles¹

¹School of Information Technology and Electrical Engineering

²School of Journalism and Communication

University of Queensland, Australia

Abstract. This paper develops a workflow for incrementally adapting visual representations, based on an existing theory of visualization creation and interpretation. The method is illustrated using an example of recurrence plots applied to text data. Incremental adaptation achieved through this workflow has the potential to provide visualization users with tools which evolve in synchronisation with their needs.

Keywords: visual analytics, adaptive analysis, visualization theory

In a wide range of domains which make use of visualization for analysis, success is dependent on being able to select an appropriate technique from a wide range of options, and also being able to evolve or tailor a visualization as the need arises. Methods which support these processes are still ineffective or inefficient. Traditionally the selection and tailoring process has been driven by the bank of available visualization methods stored either within an researcher's head, or suggested as options through a visualization system such as Spotfire (Ahlberg, 1996) or Tableau (and its predecessor Polaris) (Stolte et al., 2002)).

The field of adaptive visualization is concerned with how the display of information adapts to features of the information, the task at hand, and user characteristics (Domik and Gutkauf, 1994). Previous research in this field includes automated selection of a visualization format ((Golemati et al., 2006), (Toker et al., 2012)), adjustment of format based on user-specific perception measures (Domik and Gutkauf, 1994), and adjustment of which subsets of data should be displayed at what level of detail (Brusilovsky and Su, 2002), (Mahadev and Christie, 1996). User specified adaption has also been applied to fixed representation formats, focussing on adjustment of the data structure and level of detail (Roussinov and Ramsey, 1998). So far the range of visualization formats which can be produced using adaptive approaches has been constrained to a predetermined selection or range of formats. Here we focus on achieving incremental adaptivity in visualization: making small changes to an existing representation in such a way that it supports different analysis. This kind of adaptivity will enable a smooth transition between different kinds of analysis, and increase the extent to which new work can build on rather than supplant existing research. In this paper an existing model is used as the foundation for a workflow which deconstructs a representation and systematically generates alternatives.

The model we use was developed by Vickers, Faith and Rossiter (VFR) based on category theory and semiotics (Vickers et al., 2013) and is shown in Fig. 1. This model, which we refer to as the VFR category after its mathematical structure, decomposes visualization into intermediary objects, ranging from the system being studied to knowledge gained through visualization and transformation mappings between these objects. The VFR category has a number of features which recommend it for use in developing protocols for adaptive visual analysis. Chief among them is that it covers both the process of constructing a visualization and the reason it is created – the questions it answers and how it is interpreted. Thus intent or purpose in selecting or varying a representation can be explicitly captured in the model.

Any visualization can be deconstructed using the VFR category, and any two (or more) visualizations can be compared at each point in the category. The adaptation method we develop here is aimed at finding visualizations which are similar in the early stages of the VFR category (the left hand side of **Fig. 1**), but diverge in the later stages (right side). In other words we find new visualizations which involve only small changes from the original design, but support new kinds of analysis and interpretation.

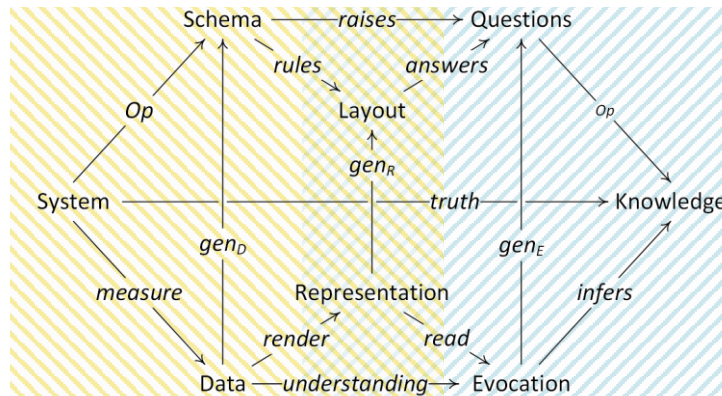


Fig. 1. The VFR category - reproduced from (Vickers et al., 2013), with shading added. Yellow shading (on the left) shows the objects and mappings involved in visualization design, while blue (right) shows those involved in interpretation of the created visualization. The representation and its generalised layout belong to both sets.

As a test case we look at the visualization technique of recurrence plots applied to language data. A recurrence plot uses a sequence of ordered data points (such as letters in a text) and shows us which pairs of positions in the sequence have the same data point values (i.e. the same character) – in other words it identifies where recurrence exists in the sequence (Webber Jr and Zbilut, 2005). We take as our original visualization a letter-based recurrence plot created from a conversation transcript used in (Angus et al., 2012). **Fig. 2.** shows a portion of the data, render and representation of this visualization decomposed into its VFR category objects. In terms of supported

analysis, this representation allows us to see where in the text letter-recurrence exists (i.e. where are the letters the same), as well as the overall level and patterns of letter recurrence.

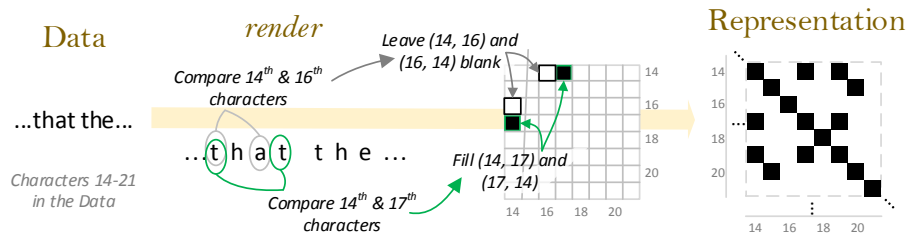


Fig. 2. A render mapping which creates a recurrence plot from an ordered list of letters applied to an example section of text

The first stage of the adaptation protocol is to find variant representations. After decomposing the existing representation, we first look for equivalent renders to the one used and examine the intermediary products and mappings they use. Related, but distinct representations can be produced by considering alternative mappings either from the data to these intermediary products or from the products to a representation. **Fig. 3** shows one example for the letter recurrence plot; the alternative render involves an intermediate product which replicates the text on each line. Letters in each line which match the character on the diagonal with a square are then replaced by squares. To create the representation in **Fig. 2**, the text is then discarded leaving only the squares (not shown in **Fig. 3**). An alternative representation is to display this intermediary product directly.

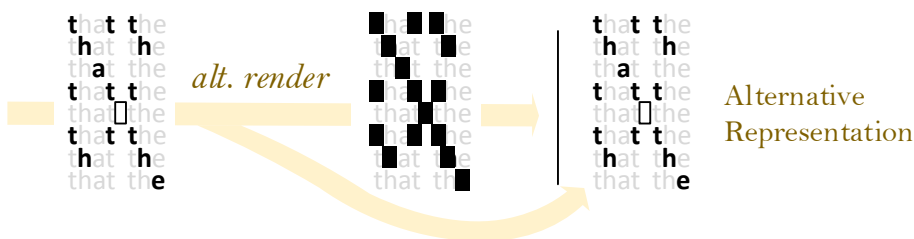


Fig. 3. An example variant render for the letter recurrence plot is shown on the left hand side of the figure. This render makes use of an intermediary product showing the text repeated and shaded to indicate recurrence. An alternative representation suggested by the render is to show the intermediary product itself (right hand side).

Having produced one or more candidate variants the next step in the method is to analyse these in terms of the right hand side of VFR category to determine the difference in the analysis supported by each variant. For the letter recurrence plot example in **Fig. 3**, the new representation allows an analyst to identify which letter or sequence of letters is recurring, information which is lost in the original representation. However, the size needed to make the letters readable means that this gain comes at the cost of cost of discovering patterns across the whole data set. This process of comparing

analysis supported allows the researcher to identify whether a variant is an adaptation in the sense that it supports different analysis, and also provides a means for choosing the most suitable of multiple possible adaptations.

The VFR category provides a foundation for building a method for incrementally adapting visualization: suited to this task because of its coverage of both visualization creation and interpretation. As shown in the example of letter recurrence plots, incrementally adaptive visualizations can be created using the VFR category to decompose an existing representation and find variations on a render mapping which support different analysis from the original. As either the data a researcher is using changes, or the research questions are refined, this method provides a means of evolving a representation to suit.

References

1. Ahlberg C (1996) Spotfire: an information exploration environment. *SIGMOD Rec.* **25**, 25-9.
2. Angus D, Watson B, Smith A, Gallois C, and Wiles J (2012) Visualising Conversation Structure across Time: Insights into Effective Doctor-Patient Consultations. *PLoS ONE* **7**, e38014.
3. Brusilovsky P and Su H-D (2002) Adaptive Visualization Component of a Distributed Web-Based Adaptive Educational System. In *Intelligent Tutoring Systems*. Cerri S, Gouardères G, and Paragauçu F (ed.), Vol. 2363, pp. 229-38. Springer Berlin Heidelberg,
4. Domik GO and Gutkauf B (1994) User modeling for adaptive visualization systems. In *Proceedings of the conference on Visualization '94.* (ed.), Vol. pp. 217-23, IEEE Computer Society Press, Washinton, D.C.
5. Golemati M, Halatsis C, Vassilakis C, Katifori A, and Lepouras G (2006) A Context-Based Adaptive Visualization Environment. In *Information Visualization, 2006. IV 2006. Tenth International Conference on.* (ed.), Vol. pp. 62-7,
6. Mahadev PM and Christie RD (1996) Minimizing user interaction in energy management systems: task adaptive visualization. *Power Systems, IEEE Transactions on* **11**, 1607-12.
7. Roussinov D and Ramsey M (1998) Information forage through adaptive visualization. In *Proceedings of the third ACM conference on Digital libraries.* (ed.), Vol. pp. 303-4, ACM, Pittsburgh, Pennsylvania, USA.
8. Stolte C, Tang D, and Hanrahan P (2002) Polaris: a system for query, analysis, and visualization of multidimensional relational databases. *Visualization and Computer Graphics, IEEE Transactions on* **8**, 52-65.
9. Toker D, Conati C, Carenini G, and Haraty M (2012) Towards Adaptive Information Visualization: On the Influence of User Characteristics. In *User Modeling, Adaptation, and Personalization*. Masthoff J, Mobasher B, Desmarais M, and Nkambou R (ed.), Vol. 7379, pp. 274-85. Springer Berlin Heidelberg,
10. Vickers P, Faith J, and Rossiter N (2013) Understanding Visualization: A Formal Approach Using Category Theory and Semiotics. *Visualization and Computer Graphics, IEEE Transactions on* **19**, 1048-61.

Work in Progress: Degree-of-Interest Based Visual Exploration of Heterogeneous Networks

Sanjay Kairam

Computer Science Department, Stanford University
Stanford, CA USA
skairam@cs.stanford.edu

Abstract. Techniques designed to support visual analysis of simple graphs may not easily accommodate networks with many node and edge types. In this paper, we discuss preliminary work on extending degree-of-interest (DOI) exploration techniques to heterogeneous networks. We first discuss requirements for computing DOI scores over multivariate network data and propose a method based on *random walks with restart*, which accommodates several challenges posed by real-world datasets. We present two complementary interfaces for visualizing the generated DOI subgraphs; through a pilot study with 12 participants, we compare these approaches and identify avenues for ongoing research.

Keywords: Graph visualization · degree-of-interest · heterogeneous networks

1 Introduction

Many datasets of interest to analysts (e.g. social interaction traces, literature collections, and biological systems) can be modeled as networks of interconnected entities of multiple types sharing various kinds of relationships. We might consider intelligence analysis as one motivating example, as it requires analyzing not only the people involved in a social network, but also the many artifacts with which these people engage, from messages to media to more abstract entities, such as discussion topics.

Top-down visual analysis of such networks can be challenging, often necessitating simplification or abstraction. PivotGraph [7] and HoneyComb [6], for instance, address this complexity by aggregating nodes along shared attributes, hiding details associated with individual actors or relationships in the data. Other approaches allow analysts to construct user-specified or computed projections of the network [2,4]. These simplified views may offer more detail, but possibly at the cost of obscuring patterns visible only when viewing many types of entities interacting simultaneously.

This paper describes our ongoing research on bottom-up visual exploration of heterogeneous networks, with a focus on presenting a subset of the data with sufficiently high resolution to support detailed analysis tasks. Specifically, we draw on prior work exploring degree-of-interest (DOI) exploration [1,3,5], extending this technique to networks with multiple types of nodes and edges. In DOI-based approaches, the analyst initiates exploration at a specific item of interest and the system returns a view of

related items. The success of these approaches lies in the strength of two primary components: a *scoring function* to compute relevance for each item with respect to the item of interest, and a *visual interface* which displays highly ranked items more prominently, allowing the analyst to focus his or her exploration on the elements of the network which may be most pertinent to the current analysis task.

2 Degree-of-Interest Scoring for Heterogeneous Networks

As a specific example to illustrate some of our objectives in developing a DOI function, consider a sample network of academic conference publication data, with nodes representing entities such as conferences, publications, authors, keywords, and edges representing various types of relationships among these nodes. For an analyst starting with the node representing the keyword “Information Visualization”, the system should identify content *relevant* to the query and of *diverse types*, presenting a mix of related papers, authors, and keywords. Finally, the system should return *structurally diverse* results, triggering opportunities for follow-up searches along various aspects of the topic of information visualization. In addition, we consider some challenges associated with real-world datasets; specifically, many datasets of interest to analysts are *large*, *continuously updated*, and may be subject to *missing or noisy data*.

Based on these objectives and our initial experiments, we propose a DOI scoring function based on *random walks with restart* (RWR), where nodes in the graph are scored according to the probability that a random walk of fixed length initiated at the query node will end there. Our implementation computes scores by iteratively simulating a large number of walks and counting the stopping points. This method is fast, easy to compute, and appropriate when datasets are large, messy, or frequently updated. Like the DOI function proposed by van Ham [5], our approach attempts to model various aspects of relevance, such as distance from the query node and a notion of a priori interest. A sharp distinction, however, is that our technique requires no information about node content, meaning that nodes can be ranked using only the network structure, even if attributes are missing. Thus, we can place and rank nodes in the graph for actors or pieces of content (such as a publication) which we know to exist, but for which we may be missing some or all of the document text or metadata.

This choice of scoring function also provides a variety of interpretable parameters that can be tuned on the fly to affect results in various ways. Changing the number of iterations, for instance, influences the tradeoff between response time and result quality. Changing the length of walks can vary the extent to which the algorithm favors nodes that are “locally relevant” to the query node vs. those which are “globally important”. Finally, attaching different weights to node and edge types can influence the probability of following particular types of paths. Computing scores interactively means that these parameters can be modified during a search session, either automatically or directly by the user. Our ongoing research efforts include identifying the most useful of these parameters and designing effective user controls which can be integrated into the interface used to present the returned contextual subgraphs.

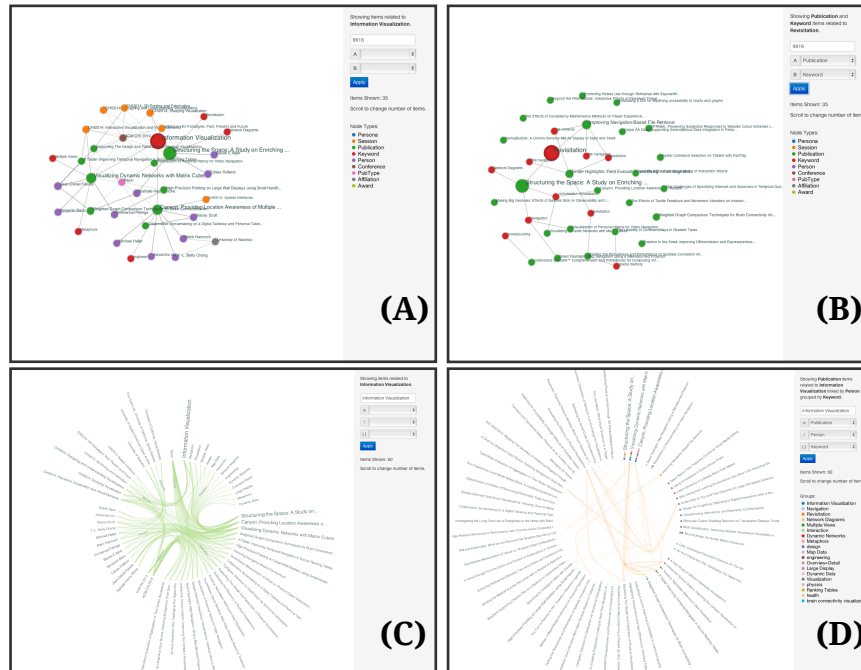


Figure 1. Figure 1a shows the initial view in the force-directed interface for a search of the keyword “Information Visualization”. Figure 1b shows the two-mode network filtered on *Publication* and *Keyword* nodes. Figure 1c shows the initial view in the radial interface for the same query. Figure 1d shows this data filtered to publications, joined by shared-author links and grouped by keyword (using colored circles).

3 Visualizing Contextual Subgraphs

Given a query node and DOI function, our system returns a subset of nodes, each with a DOI score reflecting predicted relevance to the user. Using these nodes and their connecting edges, our goal is to construct an informative, contextual diagram providing information relevant to the query node. In this section, we describe the design of two alternative prototype interfaces for constructing, presenting, and interacting with these diagrams, along with the results of a study designed to identify the strengths of each approach and elicit additional design requirements.

Force-Directed Interface. Our first prototype, based on a force-directed node-link diagram, is chosen for its familiarity and prior use in systems for DOI and heterogeneous network visualization [2,4,5]. Figure 1a shows the main view, displayed after a search has been executed. Nodes are represented by circles, with color mapped to type and size scaled by the computed DOI score. Figure 1b demonstrates how filters can be applied to support analysis questions about specific combinations of node types.

Radial Interface. The first prototype was chosen for its familiarity and simplicity of representation; the second was designed to address some issues frequently associated with node-link diagrams. First, we aim to *reduce clutter*, including the problem of overlapping node labels. Second, we hope to *facilitate path-focused tasks*; many simple analysis tasks, such as connecting co-authors, require analysis over paths of length two or greater. Finally, we aim to better *support grouping*, as even simple visual grouping can be problematic in force-directed layouts.

Figure 1c illustrates the initial view for our radial interface, designed to address some of these design concerns. Node labels are organized around a circle, grouped by type, and scaled according to DOI scores. Edges are drawn using hierarchically bundled paths. The interface enables three primary interactions. *Node filtering* allows the analyst to filter by a single node type, drawing items of this type on the screen. *Link projection* allows the analyst to choose a second type through which links are projected (e.g. showing publications which share a common author). *Grouping*, finally, allows analysts to group nodes according to a third node type (e.g. grouping publications by common keywords). Figure 1d illustrates a view filtered to *Publications* nodes, with projected shared-authorship links, grouped by *Keyword*.

Pilot Study. We evaluated both approaches through a pilot study, in which 12 participants engaged in various structured and unstructured exploration tasks over a network comprised of academic conference data; all participants were active researchers in areas covered by the data. Tasks were varied, but some examples included finding papers of interest at an upcoming conference or finding keywords relevant to papers associated with a specific author. The study utilized a within-subjects design; each participant used both the force-directed and radial interfaces, with order inverted for half the participants. Data were collected via interaction logs and subjective measures.

Subjective responses revealed roughly equal preferences for the two interfaces. As expected, the force-directed layout was judged as *simpler*, but ratings did not differ significantly on any other subjective measures. Post-test comparison questions, in which participants were asked which version they might prefer in various scenarios related to the data, revealed a similar split. Logged behavior was also similar across the two conditions. Response variables were analyzed using linear mixed-effects models, treating interface conditions and task order as fixed effects and participant ID and task ID as random effects. For both structured and unstructured tasks, no significant differences were found between interface conditions with respect to the number of distinct views, number of unique view transformations, and number of unique query nodes. One behavioral finding was that participants using the force-directed interface spent a large amount of time dragging nodes out of the way, doing so 13.6 times, on average, during unstructured tasks and 26.7 times during structured tasks.

4 Ongoing Research

We are encouraged by the fact that both interfaces were rated highly on subjective measures including *ease of use*, *simplicity*, *interestingness*, and *enjoyability*, and many participants responded casually that they would like such a tool for exploring

publication data, as no existing system offers this type of flexible, bottom-up exploration. Label legibility was still an issue in both prototypes, and our ongoing research will explore additional methods of balancing layout with the quantity of information presented to maximize the extent to which these diagrams inform the user.

Similarly, we would like to explore motivations for dragging nodes in the force-directed layout with the aim of differentiating between cases in which this behavior is adaptive (e.g. facilitating cognition, as in card-sorting) or simply reflective of frustration with the layout. While we felt that the force-directed layout was comparable in quality to many existing systems, it is likely that this interface could be improved using layout constraint techniques, such as those implemented in WebCoLa.¹

We are continuing to evaluate variants of our DOI scoring and presentation methods in order to determine how to most effectively support visual analysis over heterogeneous network data. Our expectation going into the pilot study was that participants would have a stronger subjective preference for one interface over the other; because this was not the case, the exploratory tasks chosen may not have been the most appropriate evaluation tasks, as they did not permit us to utilize quantitative measures of task “success”. In future evaluations, it may be more appropriate to choose data and tasks which more closely simulate a specific analysis scenario, allowing us to better quantify performance with respect to the quantity and quality of insights generated.

5 References

1. Furnas, G.W.: Generalized fisheye views. *ACM SIGCHI Bulletin* 17 (4), pp. 16-23 (1986)
2. Heer, J., Perer, A.: Orion: A system for modeling, transformation, and visualization of multidimensional heterogeneous networks. *Information Visualization* (2012)
3. Lee, B., Parr, S., Plaisant, C., Bederson, B.B., Veksler, V.D., Gray, W.D., & Kotfila, C.: TreePlus: Interactive exploration of networks with enhanced tree layouts. *TVCG* 12 (6), pp. 1414-26 (2006)
4. Liu, Z., Navathe, S.B., & Stasko, J.T.: Network-based visual analysis of tabular data. *VAST 2011* (2011)
5. van Ham, F. & Perer, A.: “Search, show context, expand on demand”: supporting large graph exploration with degree-of-interest. *TVCG* 15 (6), pp. 953-60. (2009)
6. Van Ham, F., Schulz, H., & Dimicco, J.: Honeycomb: Visual Analysis of Large Scale Social Networks. *INTERACT 2009*, pp. 429-442. (2009)
7. Wattenberg, M.: Visual exploration of multivariate graphs. *CHI 2006*, p. 811-818 (2006)

¹ <http://marvl.infotech.monash.edu/webcola/>

Diagrams in Patents: An exploratory introduction

Sylvan G. Rudduck¹, Mary-Anne Williams, Natalie Stoianoff.

sgrudduck@qisip.org, Mary-Anne.Williams@uts.edu.au, natalie.stoianoff@uts.edu.au
University of Technology, Sydney

Abstract.

In this submission, we present ongoing work exploring how cognitive sciences might improve the legal discipline of intellectual property. In particular, how diagrammatic representations can address problems experienced in the global patent system. In light of the problem identified in the legal discipline, our ultimate goal is to enrich patent information with rules from legal authorities and meaning from technical experts, making further provisions for the independent evaluation of resulting rules and meaning. This work takes a first step towards such a quality information system for intellectual property (QiSiP) by providing a process to collect diagrams. Discussion is made towards the common feature amongst diagrams, the difficulty of the problem domain and a potentially useful piece of existing empirical scholarship. Ongoing work is aimed at addressing the identified limitations of this work in light of results from independent evaluation.

1 Introduction

1.1 Overview

The use of visual depictions (diagrams/information graphics) as a means to communicate existing knowledge is well known [1-5]. However, little research has looked at how such external representations can facilitate the creation of new scientific knowledge [6-9]. Motivated by governmental reforms, aiming at improving the incentive systems for the creation of new knowledge [10, 11] This research is concerned with the application of knowledge representation techniques [12] to solve problems in the discipline of intellectual property law. [13]. In particular it pursues an interdisciplinary approach to representing new concepts, such as ‘shape’ [14] for practical purposes, such facilitating inventive ideas [15]. The purpose of this work is to outline a process for the collection and review of diagrams in scientific patents.

¹ Senior Graduate Researcher appreciates the past support of an Australian Postgraduate Award, the assistance of UTS-GRS and the time of the Diagram 2014 Anonymous reviewers.

2 Background

2.1 Patents and their Problems

The World Intellectual Property Organization (WIPO) defines a patent as: *‘an exclusive right granted for an invention, which is a product or a process that provides, in general, a new way of doing something, or offers a new technical solution to a problem. To get a patent, technical information about the invention must be disclosed to the public’* [16] Whilst the above definition is comprehensive, concise and lays the foundation for an intermediary means to initiate patent protection in some 140+ countries, via the Patent Co-operative Treaty (PCT). The definition does not explicitly cover the notion of patent quality.

This work continues a previous approach to defining the legal quality of a patent as its ‘correctness to rules’ [15]. A related, though more extensive approach comes from the work of Guerrini [17]. His work presented a multidisciplinary perspective to patent quality, which included a theoretical approach to measuring it along several ‘dimensions’ of Conformance, Clarity, Faithfulness, Social Utility and Commercial Success [17]².

In the US, prominent legal scholars have suggested that the underlying problem is related to the claim section of patents. In particular, the determination of the meaning [18] and communication of boundaries [19] of the what the inventor defined as their invention. The problem can be further paraphrased as *the uncertainty in the connection between words of patents and the invented thing*. [15]³ Whilst there will undoubtedly be other related or distinct problems with the patent system, this work considers – Communication (‘Notice’) and Meaning (‘words to things’) as the fundamental problems to be addressed.

2.2 Related work

There have been various dealings with issues related to, although not directly addressing, the meaning and communication of patent claims. These approaches typically reside within a single discipline, for example, within a) The legal sub-discipline of Patent law [20-22] or b) Computer science sub-discipline of information retrieval (IR) [23-25]. Discussion of legal scholarship addressing the improvement to patent quality is best left for a legal venue⁴. However, due to its technical rigor and focus on patent figures, work provided by IR community is worthy of note.

The IR consortium provided a tool to automate the extraction and searching of patent figures in bulk [23]. A component of this tool utilized the textual labels of patent figures to produce an initial hierarchy of patent figures [25] Figure labels included;

² In an independent pursuit – Q4 of our pilot survey asks the question: *‘The definition of a Quality Patent should consider: (Providing 5 related dimensions)’*. (Analysis ongoing)

³ See Paragraph 1.2 of cited reference.

⁴ For a notable exception – See Formers interdisciplinary article entitled A Psychology of IP.

Graph, Flowchart, Technical Drawing, and Photos, with various sub-labels being ordered according to Fig. 1 below.

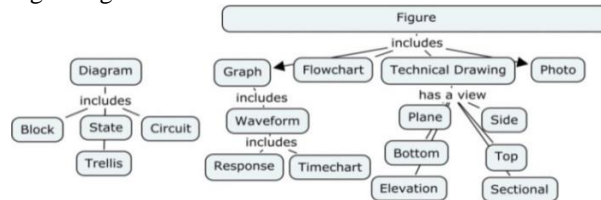


Fig. 1. Hierarchy of patent figures based on parsing of textual figure labels. Adapted from [25]

Our work seeks to make its contribution in the space in between the two disciplines of Computer Science & Patent Scholarship – That is, providing an interdisciplinary framework of how diagrams can support the claims of scientific inventions⁵.

2.3 Diagrammatic Approach

In the global legal context provided by WIPO, Rules from the PCT indicate that individuals may utilize ‘drawings...where required’ [26] But required for what?

It is further stated that ‘Drawings shall be required where they are necessary for understanding of the invention.’ (Art. 7) [27] Such a rule should be clear cut – with only a debate of ‘Understanding for what purpose?’ required. Unfortunately, the situation is immediately confused with further rules which suggest that the requirements (for drawings) vary from jurisdiction to jurisdiction. What is believed to be required is guidance through the laws and suggestions how they are related to the mind of an inventive individual – Starting first with a means collect and review the diagrams of patents.

3 Methodology

3.1 Previous work

This work tentatively follows a framework of Intermediary Theory [28]. In earlier work motivation for following this framework was ‘*due to its strong philosophical foundation and its suggested application to public policy – a key aspect of global IP law*’[15]. In light of empirical evidence on the existence of ‘shape codes⁶’ in patent databases, the same work provided a working hypothesis: ‘*that the concept of shape, embodied in visual depictions of patent figures, will provide a useful means to reduce the uncertainty between ideas (representations) of the inventor, the claims of the patent artefact and the related objects / processes in the world*’(Sect. 1.2)[15]

⁵ Including previous empirical evidence from Legal Databases, which were subjected to Independent review in the form of a survey at www.qisip.org (Under analysis)

⁶ ‘Shape codes’: ‘textual labels’ relating to conventional words for geometric entities.

In an attempt to clarify the working hypothesis, this work offers an initial manual process for collecting such visual depictions.

3.2 Overview of Collection Procedure.

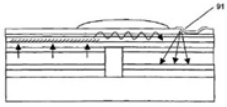
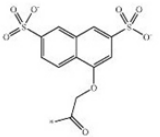
This work utilizes a previously collected sample of patents, ‘Historical Australia Nano’ (HaN). The field of Nanoscience was chosen because of its presumed importance in scientific education [29, 30] and patents [31, 32]. In hoping to bring transparency to the patent review process, the procedure is further explained below⁷:

1. An index of Raw Patent data (Patent No. / Titles herein PiD) can be obtained from the USPTO [33]⁸ and proprietary search engine being used to collect complete documents. [34]
2. Resulting patents were read and electronically annotated, to identify the text of ‘Claims’ (Clms), ‘Text figure labels’ (Tfl) and ‘Non-text figures’ (Dgram) section. Non-textual and text data was stored in a database/spreadsheet respectively.
3. Within the spreadsheet, Tfl. of patents were assigned consecutive numbers.
4. Three Tfl. were selected, via a random number generator and a further three, in a non-random manner⁹. All 6 Tfl. coming from unique Patents (PAT), which were identified with ‘###’ - the last numerals of their patent number.
5. Using the six Tfl, the corresponding Dgrams were selected and cropped for presentation.
6. Using the six identified PAT. The first claim of each was selected.
7. Noticing variation in the length of the Tfl/CLM, the first 2-5 words were presented.

4 Results

The table below shows a selection of the collected data. In addition to the visual depiction, there are three further pieces of information – Patent ID, Patent claim, and figure label – which are underlined, italic and in bold respectively.

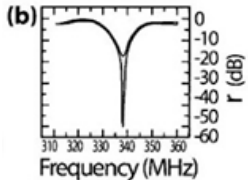
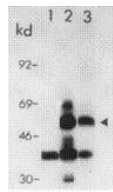
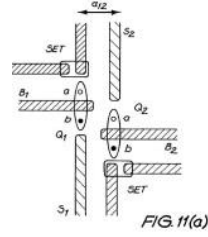
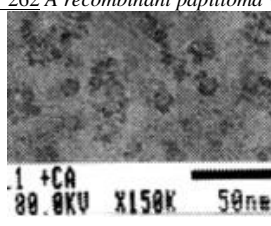
Table 1. Results from ‘Historical Australian Nano’

IP Artefact Component (Historic Australian Nano Sample)	
Randomly Selected	Non-randomly Selected
<p><u>Pat. No. `650</u> <i>A monolithically integrated biochip</i></p>  <p>Figure 9</p> <p>Fig 9,illustrates a cross-section..</p>	<p><u>Pat. No. `056</u> <i>A compound of the formula I:</i></p>  <p>IV is a formulae</p>

⁷ Broad details of its collection and descriptive statistics have been given in earlier work. See Section 3.2 Data Collection: A through C. of Reference 15.

⁸ Let Term 1=977/\$ in field ‘US Classification’ AND Term 2=AU in field Inventor Country. (Last Accessed:15/06/2014 giving several additional granted patents to HaN sample.)

⁹ Non-random selection was based on author’s curiosity in the diagrammatic forms.

<p>Pat. No. `078 A quantum device, comprising:</p>  <p>Fig 8 are measurements;</p>	<p>Pat. No. `557 A polynucleotide consisting essentially</p>  <p>Fig 2b is a western blot.</p>
<p>Pat. No. `804 A silicon integrated circuit device ..</p>  <p>Fig 11a is a plan view</p>	<p>Pat. No. `262 A recombinant papilloma virus L1</p>  <p>Fig. 2 is an electron micrograph.</p>

5 Discussion

5.1 Of Examples

Upon initial inspection, the non-expert viewer is not wrong to feel that the examples are obscure. Further, a legal practitioner may rightly note the simplification which has taken place from original patent documents. However, even without knowledge as to the technical details of invention or legal training, it is possible to notice several commonalities which may prove useful in advancing disciplinary knowledge and industry practices. A primary observation is that all the examples of diagrams contain both textual (including numerals) and non-textual elements. It is believed that such an initial primary observation, if generalizable to a larger sample or confirmed by legal authorities, may provide insights into the representation of inventive ideas.

5.2 Of Underlying problem

In focusing on the identified problem at hand - Do the presented results address of meaning and communication of claims? Not explicitly – As such issues resemble the ‘reference problem’ which has been the subject of philosophical debate for millennia (See summary at [35]). In delaying such a debate to a later date, any potential approach in the applied domain of patent law is nevertheless likely to grapple with the semantics of diagrams [36, 37].

One noteworthy approach to the joint semantics of text & diagrams is the work ‘Diagrams in the comprehension of scientific texts.’ [2] It studies the interaction of text and diagrams in the context of multiple scientific disciplines. Its teachings include

insights into taxonomy of scientific diagrams, relationships between text and diagrams, and the formation of (internal) representations. Importantly, based on the empirical results of how an expert fixes their gaze whilst reviewing text and visuals of a mechanical system, they conclude;

‘The optimal medium for communicating scientific information also depends on the skills of the reader. A diagram may be most useful if the reader has the knowledge necessary to extract the relevant information from the diagram and if the topic is sufficiently complex that the reader cannot visualize spatial representations of the information without a diagram’ (p666)[2]

Whilst the limited domain of their study, mechanical systems, may restrict the extent which their specific inferences can be applied to a broader patent domain, the conclusions that diagrams may be most useful if user has the ‘required expert knowledge’ and when ‘the topic is sufficiently complex’ - are useful starting guidelines in determining when such forms of knowledge are necessary for the understanding of patent claims.

6 Ongoing Work

Encouraged by the existence of the above (and no doubt other) analyzed work from the diagrammatic community, Future work will address the limitations of this interdisciplinary exploration. The limitation of the small sample size, whilst justified given the exploratory nature, can be increased via the formulation of a larger search query or automated collection techniques. Perhaps a more complex issue is refining the intuited approach to exploration – Such intuitions can be refined via the incorporation of decisions from legal authorities and by subjecting findings to independent evaluation, for example by a survey of senior patent examiners [38]. The results of which should shed further light on this interesting problem domain for future diagrammatic researchers.

7 References

1. Bertin, J.: *Semiology of Graphics: Diagrams, Networks and Maps*. University of Wisconsin Press (1983)
2. Hegarty, M., Carpenter, P.A., Just, M.A.: Diagrams in the comprehension of scientific texts. In: Barr, R., Kamil, M.L., Mosenthal, P., Pearson, P.D. (eds.) *Handbook of reading research*, vol. 2, pp. 641-668 (1990)
3. Olivier, P.: Diagrammatic reasoning: An artificial intelligence perspective. *Artificial Intelligence Review* 15, 63-78 (2001)
4. Larkin, J.H., Simon, H.A.: Why a diagram is (Sometimes) worth 10,000 words. *Cogn. Sci.* 11, 65-99 (1987)
5. Cleveland, W.S., McGill, R.: Graphical perception and graphical methods for analyzing scientific-data. *Science* 229, 828-833 (1985)

6. Gorman, M.E.: Introduction to Cognition in Science and Technology. *Topics in Cognitive Science* 675-685 (2009)
7. Ferguson, E.S.: The Mind's Eye: Nonverbal Thought in Technology. *Science* 197, 827-836 (1977)
8. Suwa, M., Tversky, B.: External representations contribute to the dynamic construction of ideas. In: Hegarty, M., Meyer, B., Narayanan, N.H. (eds.) *Diagrammatic Representation and Inference*, vol. 2317, pp. 341-343 (2002)
9. Cheng, P.C.H.: Scientific discovery with law-encoding diagrams. *Creativity Research Journal* 9, 145-162 (1996)
10. Hargreaves, I.: *Digital Opportunity: an Independent Review of IP and Growth*. (2011)
11. Cuttler, T.: *Venturous Australia: Building Strength in Innovation*. (2008)
12. Williams, M.A., McCarthy, J., Gärdenfors, P., Stanton, C., Karol, A.: A grounding framework. *Auton Agent Multi-agent system* 19, 272-296 (2009)
13. Reynolds, R., Stoianoff, N.: *Intellectual Property: Text and essential cases*. Federation Press (2012)
14. Rudduck, S.G., Williams, M.A.: Conceptual ternary diagrams for shape perception: a preliminary approach. In: Bertel, S. (ed.) *AAAI 2010 Spring Symposium*, Stanford, CA (2010)
15. Rudduck, S.G., Williams, M.A., Stoianoff, N.: Visualizing the 'Shape of Quality': An Application in the Context of Intellectual Property. In: *Shapes 1.0 The shape of things*. CEUR, (2011)
16. Patent Scope <http://www.wipo.int/patentscope/en/>
17. Guerrini, C.J.: Defining Patent Quality. *Fordham L. Rev.* Forthcoming 82, (2013)
18. Burk, D.L., Lemley, M.A.: *The Patent Crisis and How the Courts can solve it*. University of Chicago Press (2009)
19. Bessen, J., Meurer, M.: *Patent Failure: How judges, Bureaucrats, and Lawyers put innovators at risk*. Princeton University Press (2008)
20. Feldman, R.: Plain Language Patents. *Texas Intellectual Property Law Journal* 17, 298-304 (2009)
21. Fromer, J.C.: *A Psychology of Intellectual Property*. Northwestern University Law Review 104, (2010)
22. Noveck, B.S.: Peer to Patent: Collective Intelligence, Open Review and Patent Reform. *Harvard Journal of Law & Technology* 20, 123-162 (2006)
23. Bhatti, N., Hanbury, A.: Image search in patents: a review. *International Journal on Document Analysis and Recognition* 16, 309-329 (2013)
24. Vrochidis, S., Papadopoulos, S., Moutzidou, A., Sidiropoulos, P., Pianta, E., Kompatsiaris, I.: Towards content-based patent image retrieval: A framework perspective. *World Patent Information* 32, 94-106 (2010)
25. Wanner, L., Baeza-Yates, R., Brüggmann, S., Codina, J., Diallo, B., Escorsa, E., Giereth, M., Kompatsiaris, Y., Papadopoulos, S., Pianta, E., Piella, G., Puhlmann, I., Rao, G., Rotard, M., Schoester, P., Serafini, L., Zervaki, V.: Towards content-oriented patent document processing. *World Patent Information* 30, 21-33 (2008)
26. WIPO: PCT-Article 3. In: United Nations, U. (ed.) (2002)
27. WIPO: PCT-Article 7. In: United Nations, U. (ed.), (2002)

28. Shields, P.M., Tajalli, H.: Intermediate Theory: The missing link to successful student scholarship. Faculty Publications - Political Science 39, (2006)
29. Shapter, J., Ford, M.J., Maddox, L.M., Waclawik, E.R.: Teaching Undergraduates Nanotechnology. International Journal of Engineering Education 18, 512-518 (2002)
30. Tang, K.-S.: Instantiation of multimodal semiotic systems in science classroom discourse. Language Sciences 37, 22-35 (2013)
31. USPTO: Class 977 Nanoscience In: Support, O.o.C. (ed.), (2011)
32. Lemley, M.A.: Patenting Nanotechnology. Stanford Law Review 58, (2005)
33. USPTO, Patent Database (2010) <http://patft.uspto.gov/netahtml/PTO/search-bool.html>
34. Google Patents <https://google.com/patents>
35. Reimer, M.: Reference. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy (2010)
36. Tufte, E.: Micro/Macro Readings. Envisioning Information, (1990)
37. Hegarty, M.: Diagrams in the mind and in the world: Relations between internal and external visualizations. In: 3rd International Conference on Diagrams, pp. 1-13. (2004)
38. Rudduck, S.G.: Online Survey of Senior Patent Examiners. UTS-IP Australia, Sydney. (2014)

Hunches and Sketches: rapid interactive exploration of large datasets through approximate visualisations

Advait Sarkar, Alan F Blackwell, Mateja Jamnik, Martin Spott

Computer Laboratory, University of Cambridge, UK
BT Research and Technology, Ipswich, UK
{advait.sarkar, alan.blackwell, mateja.jamnik}@cl.cam.ac.uk
martin.spott@bt.com

Abstract. Information visualisation presents powerful techniques for data analytics. However, rendering visualisations of big datasets is impractical on commodity hardware. There is increasing interest in approaches where data sampling and probabilistic algorithms are used to support faster processing of large datasets. This approach to approximate computation has not yet paid close attention to the way that approximate visualisations are perceived and employed by human users, as a specific variety of diagrammatic convention. Our intent is to apply this understanding of approximate visualisations as a diagrammatic class to mainstream data science and information visualisation research.

Keywords: visualising uncertainty, information visualisation, exploratory data analysis, approximate inference, big data, sketches.

1 Visual analysis, large datasets, and uncertainty

The utility of visualisation for data analysis cannot be understated. The power of the human perceptual system paired with the visualisation capabilities of modern software tools allows for the rapid detection of trends, outliers, and comparisons of quantities – even by those without statistical expertise. Visualisations are also useful for those with deeper analytical skill. The space of analytical questions one can ask of a particular dataset is infinite, but only some questions yield interesting answers. Consequently, exploratory data analysis is divided between two approaches: the “top-down”, hypothesis-testing approach wherein a specific statistical technique is used to answer a specific statistical question (e.g. *Is there a significant difference between these groups?*, or *How does a change in X affect Y ?*), and the “bottom-up”, hypothesis-generation approach where the interesting questions are identified and formulated (e.g. *Should I investigate the relationship between variables X and Y ?*).

Cognitive task analysis of this sensemaking process suggests that experienced analysts often invoke the two processes in an opportunistic mix [1]. Visualisations can help rapidly prune the space of interesting hypotheses, and it can help verify

many of these hypotheses [2], which spares the analyst the effort of conducting a more elaborate statistical investigation of a question that in hindsight turns out to be uninteresting, or the wrong question to ask.

Shneiderman claims that a combined approach would enable more effective exploration whilst giving users a greater sense of control over the direction their exploration takes [3]. Bertini and Lalanne call for researchers to identify which aspects of analytical problems can be best solved using the human perceptual system, which are best solved using machine learning techniques, and then design for this blend of strengths [4]. Keim et al. refer to tools which embrace the idea of human-machine collaboration to solve analytical problems as “advanced visual analytics interfaces” [5].

There are many such advocates of increased integration between sophisticated statistical techniques and information visualisation tools. While this is an attractive idea, recent increases in the sheer volume of data can make visual techniques inaccessible to those attempting to perform analysis on commodity hardware. For instance, a scatterplot of 10,000 data points renders relatively quickly in Microsoft Excel or R on a commodity desktop computer. However, as of this writing it is grindingly slow to render a scatterplot of 10,000,000 points. This is the situation we often find ourselves in today. It is not conducive at all to rapid interactive exploration, and defeats the benefits of visualisation. This problem is unlikely to be alleviated by advances in hardware, as the growth in data volumes is facilitated in part by improved processing capacities. Advances in distributed computing are similarly a double-edged sword, potentially improving the computing power available for rendering visualisations but also facilitating data volume growth.

One solution to this problem is to not interact with the entire dataset, but to first reduce or transform it. For instance, a small representative sample could help generate/eliminate many of the same candidate hypotheses as if one were operating on the entire dataset. Besides sampling [7], a number of approximation techniques have been developed in the past few decades that allow for fast processing of large datasets in exchange for small, quantifiable error bounds, including *sketches* and *online aggregation* [8, 9]. These advances have led to the development of database tools that can perform fast approximate queries [10].

An important note about terminology: the aforementioned “sketches” are in fact simply data structures and algorithms. They are only sketches in the sense that they are approximations of the original dataset; they are otherwise unrelated to the normal use of the word “sketch”, i.e. they are not intrinsically visual entities. For instance, the Bloom filter [11] is a data structure that represents a mathematical set and supports fast approximate membership querying. It does not represent the set exactly, but rather hashes items into a compact bit vector that approximates, or sketches, the original set. While such techniques do not use the word “sketch” in more than a metaphorical manner, the idea that these approaches could be augmented with visualisations appears to be an interesting avenue for exploration.

2 From sketches of large datasets to hunches

Our first main proposition is that data summarisation techniques can be used to interactively render approximate, exploratory visualisations of large datasets. For instance, in Figure 1, the plots on the left are of relatively large datasets. They render slowly and are therefore difficult to interact with. The plots on the right use samples or sketches of those datasets, and render much faster.

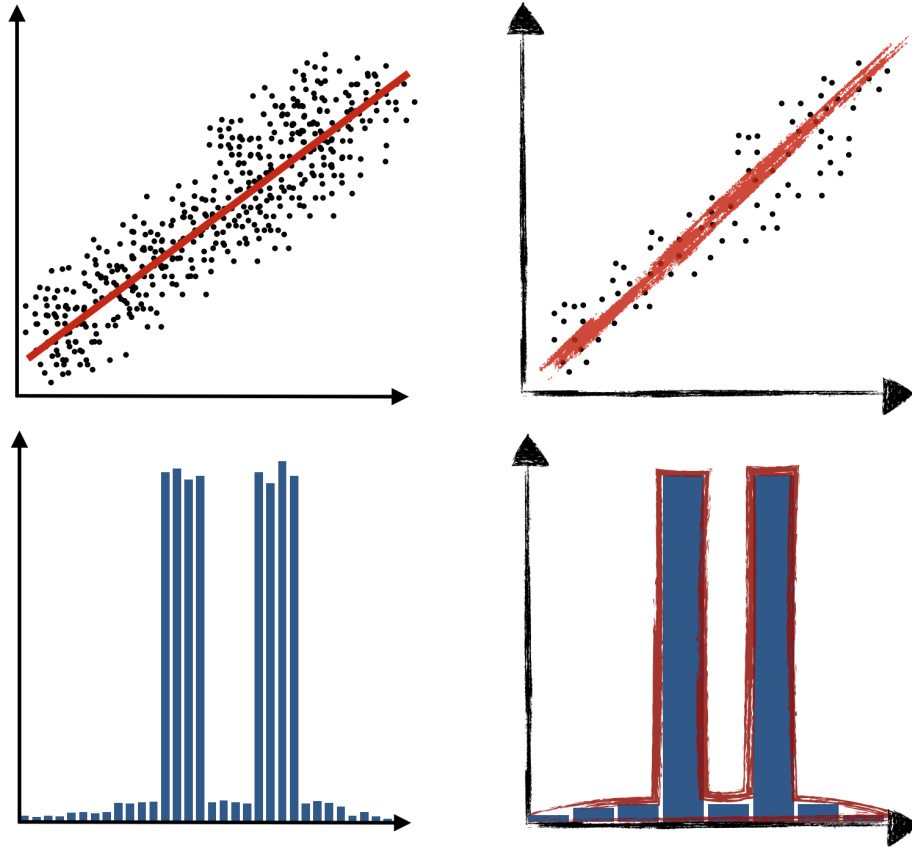


Fig. 1. Exact and approximate visualisations.

Our second main proposition is that depicting summaries of large datasets is a potentially useful application for techniques for visualising uncertainty. There are many such techniques, perhaps the most familiar being the use of error bars in bar charts and histograms. The literature discusses a variety of other techniques [12–14], including the use of transparency, blurring, painterly rendering [15, 16], and animation [17]. In particular the use of informal, sketch-like visualisations is thought to influence willingness to interact with and question the visualisation [18]. As noted by Eckert et al. [19], sketches are not simply degraded versions of a canonically accurate visual representation, but support specific cognitive and social functions.

Visualisations of uncertainty emphasise that these summaries will not support exact inference, but instead facilitate rapid informal reasoning and the formation of “hunches” – approximate hypotheses and heuristics for exploring the hypothesis space. Hunch-driven reasoning yields informal answers to open-ended questions an analyst might have (e.g. *Does this look like signal or noise? Does there appear to be cluster structure in the data? What is the general shape of the distribution? Is there an inflection point in the time series?*) before formulating specific statistical questions. These hunches may be produced on a mixed-initiative basis, i.e. collaboratively by the user and the system, thus providing a new interaction metaphor for “intelligent discovery assistants” [20].

The upper right graph in Figure 1 shows a reduced dataset which is much faster to render than the full dataset to its left. While the slope of the trend line may differ from the true slope of the trend line for the entire dataset, and the confidence intervals of any regression analysis might be wider, the reduced dataset is sufficient for the analyst to form the hunch (or informal hypothesis) of a linear relationship. The approximate nature of this hypothesis is expressed through its informal rendering, emphasising that it is not the regression coefficients that are important, but rather that a linear model may be viable. Similarly, the histogram in the lower right may have been created using a fast approximate cardinality estimator such as the linear counting algorithm [21]. It is an imperfect representation of the dataset to its left, however, the important observation is that a bimodal distribution exists, not the specific frequencies being represented.

Going forward, it will be important to study and identify several common types of these visual insights. While it would be worthwhile to demonstrate that certain transformations of the original dataset through sketching and sampling techniques will necessarily preserve these insights, it is also important to consider how we might visualise transformations that make no such guarantees or have probabilistic error bounds, which would greatly expand the range of techniques available for these interactive visualisations.

3 Conclusion

We have presented a vision for a programme of research into new tools for the interactive analysis of large datasets through approximate visualisations. These combine fast approximation techniques and techniques for visualising uncertainty, yielding new approaches to interacting with approximate visual hypotheses, or “hunches”. These approaches have the potential to afford rapid interaction with large datasets through conventional, accessible modern tools for information visualisation, running on commodity hardware.

Acknowledgements

Advait Sarkar is funded through an EPSRC Industrial CASE studentship sponsored by BT Research and Technology, and also through a Premium Studentship from the University of Cambridge Computer Laboratory.

References

1. Pirolli, P., Card, S.: The Sensemaking Process and Leverage Points for Analyst Technology as Identified Through Cognitive Task Analysis. *Proceedings of International Conference on Intelligence Analysis*, 5, 2-4 (2005).
2. Keim, D.A.: Visual Exploration of Large Data Sets. *CACM* 44(8), pp. 38-44 (2001)
3. Shneiderman, B.: Inventing discovery tools: combining information visualization with data mining. *Information Visualization*, vol. 1, no. 1, pp. 5-12, (2002)
4. Bertini, E., Lalanne, D.: Investigating and reflecting on the integration of automatic data analysis and visualization in knowledge discovery. *ACM SIGKDD Explorations Newsletter*, 11(2), 9 (2010)
5. Keim, D. A., Bak, P., Bertini, E., Oelke, D., Spretke, D., Ziegler, H.: Advanced visual analytics interfaces. *Proc. AVI '10*, 3 (2010)
6. Blackwell, A. F., Church, L., Plimmer, B. Gray, D.: Formality in Sketches and Visual Representation: Some Informal Reflections. *VLHCC workshop* 11-18 (2008)
7. Chaudhuri, S., Das, G., Narasayya, V.: Optimized stratified sampling for approximate query processing. *ACM Transactions on Database Systems*, 32(2), 9 (2007)
8. Cormode, G.: Sketch techniques for massive data. *Synposes for Massive Data: Samples, Histograms, Wavelets and Sketches*, 1-3 (2011)
9. Hellerstein, J. M., Haas, P. J., Wang, H. J.: Online aggregation. *ACM SIGMOD Record*, 26(2), 171-182 (1997)
10. Agarwal, S., Mozafari, B., Panda, A., Milner, H., Madden, S., Stoica, I.: BlinkDB: queries with bounded errors and bounded response times on very large data. *Proc. 8th ACM European Conference on Computer Systems* (pp. 29-42) (2013)
11. Bloom, B. H.: Space/time trade-offs in hash coding with allowable errors. *Communications of the ACM*, 13(7), 422-426 (1970)
12. Johnson, C. R., Sanderson, A.: A next step: Visualizing errors and uncertainty. *Computer Graphics and Applications*, IEEE, 23(5), 6-10 (2003).
13. Zuk, T., Carpendale, S.: Theoretical analysis of uncertainty visualizations. *Proc. SPIE 6060, Visualization and Data Analysis 2006*, 606007 (2006)
14. Thomson, J., Hetzler, E., MacEachren, A., Gahegan, M., Pavel, M.: A typology for visualizing uncertainty. In *Electronic Imaging 2005* (pp. 146-157). *International Society for Optics and Photonics* (2005)
15. Boukhelifa, N., Bezerianos, A., Isenberg, T., Fekete, J.: Evaluating sketchiness as a visual variable for the depiction of qualitative uncertainty. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 18(12), 2769-2778 (2012)
16. Wood, J., Isenberg, P., Isenberg, T., Dykes, J., Boukhelifa, N., Slingsby, A.: Sketchy rendering for information visualization. *IEEE TVCG*, 18(12), 2749-2758 (2012)
17. Ehlschlaeger, C. R., Shortridge, A. M., Goodchild, M. F.: Visualizing spatial data uncertainty using animation. *Computers & Geosciences*, 23(4), 387-395 (1997)
18. Bresciani, S., Blackwell, A. F., Eppler, M.: A Collaborative Dimensions Framework: Understanding the mediating role of conceptual visualizations in collaborative knowledge work. *Proc. 41st HICSS* (pp. 364-364). *IEEE* (2008)
19. Eckert, C., Blackwell, A., Stacey, M., Earl, C., Church, L.: Sketching across design domains: Roles and formalities. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, 26(3), 245-266 (2012)
20. Serban, F., Vanschoren, J., Kietz, J.-U., Bernstein, A.: A survey of intelligent assistants for data analysis. *ACM Computing Surveys*, 45(3), 1-35. (2013)
21. Whang, K. Y., Vander-Zanden, B. T., Taylor, H. M.: A linear-time probabilistic counting algorithm for database applications. *ACM TODS*, 15(2), 208-229 (1990)

Drawing Euler Diagrams and Graphs in Combination

Mithileysh Sathiyarayanan, *Student Member, IEEE*

University of Brighton, UK
M.Sathiyarayanan@brighton.ac.uk

Abstract. Euler diagrams are an attractive information visualization tool largely used in many application areas such as medicine and engineering. Graphs are also visualization tool widely used to visualize large amounts of interconnecting data in diverse application areas such as ontology modelling, bioinformatics and social network analysis. There are existing methods which combine both Euler diagrams and graphs with limited results (sub-optimal layouts are produced). Our main aim of this work is to significantly improve the analysis of grouped network data using Euler diagrams with graphs by automated visualization. This will allow the user or data analyst to navigate easily through large sets of curves along with graphs by automatically producing effective visualizations (optimal layouts).

1 Introduction

Euler diagrams represent sets and graphs are often used to represent networked data (i.e. items and their relationships). They can be used in combination and have the potential to be a powerful technique for visualizing and analyzing large and complex data sets. For example, in figure 1 the Euler diagram on the left visualizes the sets of users of the social networks twitter, google+, orkut and facebook. The graph in the middle uses edges to depict connection relationships e.g. twitter followers, facebook friends and so on. Lastly, the diagram on the right combines the two visualizations. Euler diagrams don't usually have constants but they are quite often added for convenience.

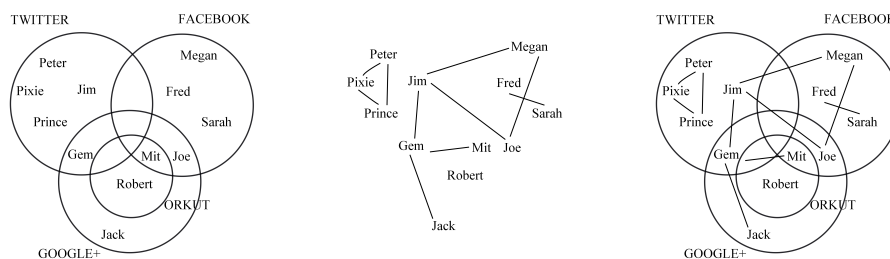


Fig. 1. An Euler diagram, a graph and the two in combination

A real world application for this visualization is in the context of advertising revenue: a company that sells targeted advertising space can charge more for displaying promotional material to highly connected, multiple network users, than to more isolated users. In our example, the combined visualization readily shows that Gem and Jim both have the most connections, and Gem uses more types of social media. Thus, advertising to Gem and Jim has the potential to reach other users more quickly, if they share the link. Moreover, advertising targeted at Gem has the immediate potential to proliferate across more social networking sites than is the case for Jim. By contrast, advertising to Peter, Pixie, and Prince is of less value because they have few connections and none outside twitter. Using the two visualizations separately does not afford the same immediacy of information extraction.

Existing attempts to automatically layout graphs (items connected by lines) and closed curves in combination to visualize groupings in networked data have produced somewhat limited results [1], [2], [3], [4], [5]: visual tools for supporting the interrogation of such data are seriously lacking. In particular, data analysts are currently not supported when they wish to visualize grouped network items. It is this problem that we will address. The main aim of this research is to allow the user or data analyst to navigate through large sets of curves (Euler diagrams) along with numerous data items (graphs) easily by automatically producing visualizations. Since socio-technical systems have grown in complexity, they have become increasingly difficult for humans to navigate and understanding the relationships between the data items is, thus, hard. The combination of Euler diagrams and graphs gives the relationship between the sets and the data items which will help people visualize, analyze and tailor large socio-technical systems. We will be developing Euler diagrams and graphs in combination and this includes devising new theory alongside practical software. Our approach will be to determine whether the state-of-the-art drawing methods for Euler diagrams and for graphs can be merged and extended. This will include the development of novel drawing methods and layout tools.

2 Automatically drawing diagrams

We now briefly discuss approaches to automatically drawing Euler diagrams and automatically drawing graphs and describe how they might be combined.

Euler diagram drawing methods: these methods start with *abstract description* of the required diagram. These descriptions specify the set intersections to be visualized. An important consideration are the *well-formedness properties* possessed by the diagrams. Examples of such properties are that the curves do not self-intersect and that there are no triple points of intersection. Diagrams that are not well-formed are considered to reduce user comprehension. An interesting aspect of Euler diagrams is that some of them cannot be drawn without breaking one or more well-formedness properties. There are three classes of drawing methods which attempt to draw well-formed Euler diagrams where

possible. These classes are *Dual Graph*, *Inductive* and, of particular interest to us, *Circle-Based* drawing methods.

The circle-based drawing method was recently devised by Stapleton et al. [6]. The main contribution of this method is to automatically draw an Euler diagram using only circles to represent any data set where the data are classified into categories. This drawing process uses strategies to transform the abstract description. These transformations are necessary because not all abstract descriptions can be drawn with circles. Whilst layout improvement is certainly possible, the diagrams achieved using this approach ensure nearly all of the well-formedness properties are possessed. The circle-based method is an ideal candidate for extending because it produces effective layouts and can always draw a diagram to represent the given data.

Graph drawing methods: these methods start with a set of vertices and a set of edges that connect the vertices, analogous to the abstract descriptions of Euler diagrams. Again, it is important to consider the *aesthetic properties* possessed by graphs. One such property is that the edges do not cross. Many algorithms take into account aesthetic properties and have been devised over the last 30 years. These algorithms can be categorized as follows: *force-directed*, *dimension reduction* and *multi-level* layout methods [7].

Force-directed techniques remain popular because of their simple form and they can be easily implemented in code. An example for generating graphs using force-directed methods is *CCVisu*, where brushing vertices are repulsed. An advantage of force-directed methods is that they yield good quality results for graphs with up to 50-100 vertices. Some of the disadvantages are complex and non-optimal layouts when drawing graphs with above 100 vertices and high running time. The force-directed method is a possible choice for extending because, for example, forces can be used to repulse vertices from Euler diagrams' curves.

Combined drawing methods: By contrast to the notations used separately, there is not a well-developed theory for drawing Euler diagrams and graphs in combination. An approach to solving our research problem is to draw the Euler diagram then the graph [1], or vice versa. However, this leads to sub-optimal diagrams. For example, in figure 2 (a), the Euler diagram was drawn first, and this resulted in the graph having edge crossings. Likewise, in figure 2

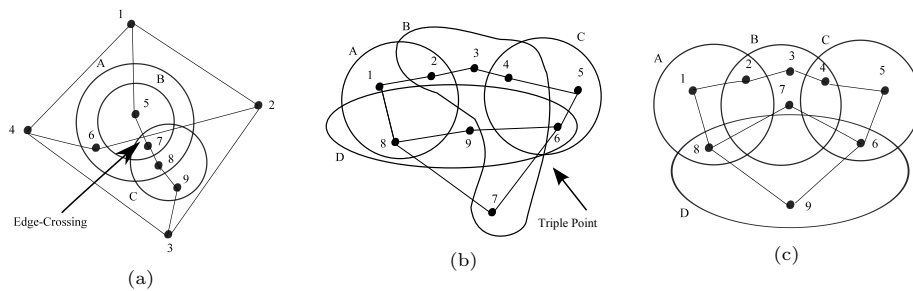


Fig. 2. Illustrating Euler diagram along with a graph: well-formedness

(b), the graph was drawn first and this resulted in a triple point, yielding a non-well-formed Euler diagram. By contrast, in figure 2 (c), which represents the same data as figure 2 (b), the layouts of the Euler diagram and the graph are both well-formed and thus not compromised. Therefore, we need methods that take account of both the Euler diagram and the graph, in combination, when constructing diagrams.

As with the two individual notations, we want to produce well-formed diagrams. We inherit well-formedness properties from each notation and have defined new ones: no curve and vertex duplicated labels, no concurrency between curves and edges, no n-points between curves and edges, no brushing points between diagrammatic elements, no edges disconnecting zones and no edges disconnecting basic regions. These well-formedness properties will help us develop good layout algorithms and drawing methods for Euler diagrams and graphs.

3 Discussion and Future Work

The aim of the research is to significantly improve the analysis of grouped network data using Euler diagrams with graphs by automated visualization. Ultimately we want to find an extension of the circle-based and the force-directed methods for Euler diagrams along with graphs. This will require diagram descriptions and the associated definitions to be extended to the combined context. Secondly, we will develop software that extends existing visualization tools and allows access to the techniques developed in the research. Finally, we will evaluate the effectiveness of the layouts produced and identify required improvements. In particular, we will conduct empirical studies and use the results to improve our novel layout techniques so that they produce better final diagrams.

References

1. P. Rodgers, P. Mutton, and J. Flower, "Dynamic Euler diagram drawing," in *IEEE Symp. on Visual Languages and Human-Centric Computing*, 2004, pp. 147–156.
2. P. Simonetto, D. Auber, and D. Archambault, "Fully automatic visualisation of overlapping sets," *Computer Graphics Forum*, vol. 28, no. 3, 2009.
3. N. Riche and T. Dwyer, "Untangling Euler diagrams," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 6, pp. 1090–1099, 2010.
4. W. Meulemans, N. Riche, B. Speckmann, B. Alper, and T. Dwyer, "Kelfusion: A hybrid set visualization technique," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 11, pp. 1846–1858, Nov 2013.
5. C. Collins, G. Penn, and S. Carpendale, "Bubble sets: Revealing set relations with isocontours over existing visualizations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 6, pp. 1009–1016, 2009.
6. G. Stapleton, J. Flower, P. Rodgers, and J. Howse, "Automatically drawing Euler diagrams with circles," *Journal of Visual Languages and Computing*, vol. 23 (3), pp. 163–193, 2012.
7. H. Gibson, J. Faith, and P. Vickers, "A survey of two-dimensional graph layout techniques for information visualisation," *Information Visualization*, vol. 12, pp. 324–357, 2012.

Storyline Visualization: Aesthetics and Legibility from Observing Art

Yuzuru Tanahashi and Kwan-Liu Ma

University of California Davis, VIDI Research Group,
One Shields Avenue Davis, California, USA
`yatanahashi@ucdavis.edu, ma@cs.ucdavis.edu`
<http://vidi.cs.ucdavis.edu/>

Abstract. Storyline visualization is a technique that effectively portrays both global trends and detailed transitions of evolving group formations in time-varying social networks. This technique was inspired by a hand-drawn illustration in XKCDs “Movie Narrative Charts” [5]. This paper presents the steps we took to transform a concept introduced by hand-drawn illustrations into a computationally automated data visualization technique that retains the aesthetics and legibility achieved by the professional artists.

Keywords: storyline visualization, layout algorithm, time-varying graph.

1 Introduction

Storyline visualization is a simple and elegant technique for portraying the temporally dynamic changes in community structures in social networks. Properly constructed storyline visualizations can convey both global trends and local interactions in the data in a single picture. This visualization technique was inspired by Munroe’s XKCD webcomic “Movie Narrative Charts” [5] depicting a visual summary of a series of movie plots (see Figure 1). In this hand-drawn illustration, each line represents a movie character. These lines converge and diverge as the characters interact and separate through the story’s plot.

Many visualization researchers have explored techniques for incorporating similar visual metaphors into various visual analytics tools [1, 3, 6, 7]. However, these early efforts focused on domain-specific visualizations and were based on a set of rudimentary design guidelines. Therefore, these visualization techniques either were overly simplified or could not achieve the aesthetics and legibility comparable to the works of professional artists. This paper presents an overview of our efforts [8] in transforming the concept introduced by a hand-drawn illustration into a computationally automated data visualization technique.

2 Steps to Storyline Visualization

There are three steps in transforming the visual concept based on a set of hand-drawn illustrations into an automated storyline visualization technique. The first

2 Storyline Visualizations: Aesthetics and Legibility from Observing Art

step is the interpretation of the illustrations. In this step, we extract the design guidelines from the original hand-drawn illustrations by observing the image and interpreting the artist’s intentions. The second step is the deconstruction of the illustrations. In this step, we deconstruct the image into simple visual units that can be easily expressed by a simple data model. The third step is the formulation of the generation process. In this step, we formulate the computational process for generating legible storyline visualization layouts using the design guidelines and the data model.

2.1 Design Guidelines

Figure 1 (Top) shows a part of the original hand-drawn illustrations made by Monroe [5]. After an extensive observation of a series of similar illustrations, we have extracted three design principles: 1) Lines representing interacting characters must be adjacent; 2) Otherwise, lines must not be adjacent; and 3) A line must not deviate unless it converges or diverges with another line. These three design principles allow us to define the underlying structure of storyline visualization, which then helps us deconstruct the illustration into visual units.

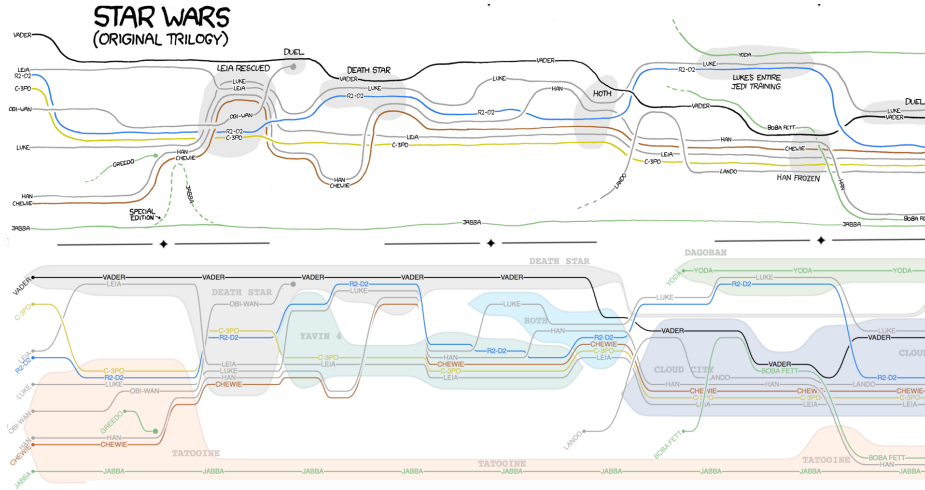


Fig. 1. (Top) An excerpt from [5] depicting the first half of the *Star Wars* movie plot. The gray background highlight the important scenes in the movie. (Bottom) A storyline visualization of the same movie generated by our layout algorithm. The background colors indicate the different planets on which the characters reside.

In addition to these design principles, we define three aesthetic dimensions that allow us to measure the aesthetics and legibility of the visualization. The ideal storyline layout contains: 1) Minimum line wiggles; 2) Minimum line crossings; and 3) Minimum white space gaps. These metrics are based on previous research that discusses the legibility and aesthetics of visualizations [2, 9].

2.2 Data Model

We designed a data model called *interaction sessions* by deconstructing a storyline visualization into a series of blocks containing bundled lines. Each interaction session represents a time period of a specific group formation and consists of three properties: start time, duration, and members. Based on the three design principles discussed above, each interaction session visually translates into a rectangular block containing a set of horizontal lines whose x-position, width, and height corresponds to the start time, duration, and the number of members.

2.3 Layout Algorithm

Using the interaction session data model, we construct a three-step process for computationally generating storyline visualizations from data. The first step is determining the topological layout of the blocks representing the interaction sessions. The second step is rearranging the lines within each block. The third step is compressing the layout by removing excessive gaps (white space) in the layout. Figure 2 shows a flowchart depicting our layout algorithm based on Genetic Algorithm.

Each genome encodes a topological layout of the interaction session blocks. Once the topological layout of the blocks is extracted from the genome, the line rearranging process adjusts the order and alignment of each block's internal lines to minimize line deviations and crossings. Finally, the storyline layout is compressed in a space efficient layout by removing the excessive white space between blocks. A detailed description of these procedures can be found in [8], and some alternative approaches have also been discussed in [4].

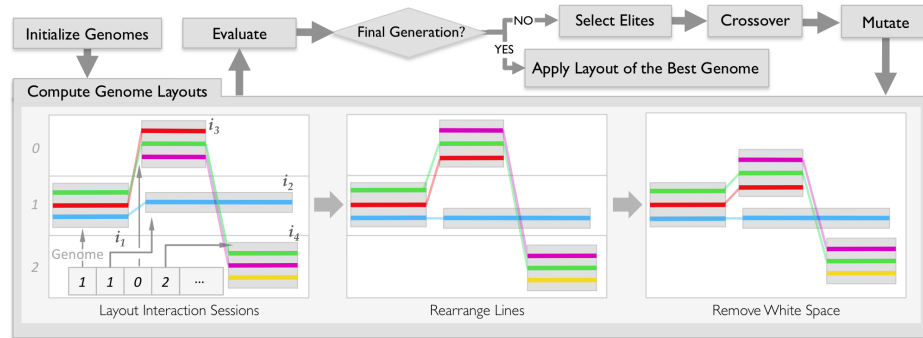


Fig. 2. A flowchart of the layout algorithm for generating legible storyline visualizations. The foundation of this algorithm is based on Genetic Algorithm with the objective of optimizing the layout with respect to the three aesthetic dimensions.

3 Discussion

The ability to computationally generate legible storyline visualizations from data not only facilitates real-world applications, but also allows further innovation to the technique. Figure 1 (Bottom) and 3 show examples of storyline visualization

generated by our layout algorithm. In these examples, the storyline visualization is extended to incorporate spatial information in addition to the temporal changes in group formation. In recent efforts, we have also been developing a framework for generating legible storyline visualizations from streaming data extending its applications to real-time information analyses. As future work, we plan to explore possibilities in developing advanced abstraction techniques for large-scale storyline visualizations. For example, integrating storyline visualizations with other visualization techniques such as stream graph will be able to provide analysts with a simple view for showing global trends while maintaining detailed views for important parts of the visualization. We are also currently developing new interaction techniques to help analysts explore and navigate large-scale storyline visualizations.

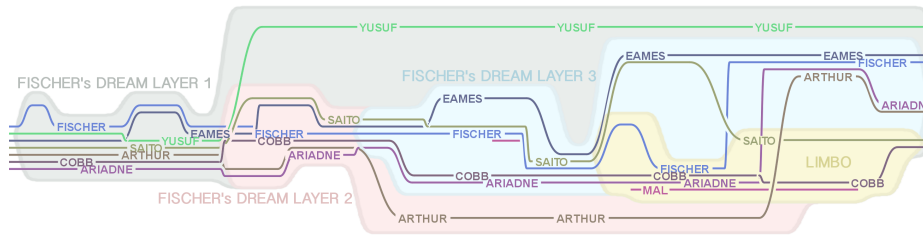


Fig. 3. Part of a storyline visualization depicting the ending of the movie *Inception*. Background colors indicate the different dream-worlds in which the characters dwell.

References

1. W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu, and X. Tong. *IEEE Transactions on Visualization and Computer Graphics*.
2. J. Díaz, J. Petit, and M. Serna. A survey of graph layout problems. *ACM Comput. Surv.*, 34:313–356, 2002.
3. N. W. Kim, S. K. Card, and J. Heer. Tracing genealogical data with timenets. In *Proceedings of the International Conference on Advanced Visual Interfaces*, pages 241–248, New York, USA, 2010. ACM.
4. S. Liu, Y. Wu, E. Wei, M. Liu, and Y. Liu. Storyflow: Tracking the evolution of stories. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2436–2445, 2013.
5. R. Munroe. Xkcd #657: Movie narrative charts. <http://xkcd.com/657>, 2009.
6. M. Ogawa and K.-L. Ma. Software evolution storylines. In *Proceedings of the 5th international symposium on Software visualization*, pages 35–42, New York, USA, 2010. ACM.
7. K. Reda, C. Tantipathananandh, A. Johnson, J. Leigh, and T. Berger-Wolf. Visualizing the evolution of community structures in dynamic social networks. *Computer Graphics Forum*, 30(3):1061–1070, 2011.
8. Y. Tanahashi and K.-L. Ma. Design considerations for optimizing storyline visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2679–2688, 2012.
9. C. Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2004.

Index of Authors

—/ **A** /—
Angus, Daniel 1

—/ **B** /—
Blackwell, Alan F. 18
Byrne, Lydia 1

—/ **J** /—
Jamnik, Mateja 18

—/ **K** /—
Kairam, Sanjay 5

—/ **M** /—
Ma, Kwan-Liu 27

—/ **R** /—
Rudduck, Sylvan G. 10

—/ **S** /—
Sarkar, Advait 18
Sathiyarayanan, Mithileysh 23
Spott, Martin 18
Stoianoff, Natalie 10

—/ **T** /—
Tanahashi, Yuzuru 27

—/ **W** /—
Wiles, Janet 1
Williams, Mary-Anne 10