APPENDICES

Design Space of Origin-Destination Data Visualization

Martijn Tennekes, Statistics Netherlands, the Netherlands
Min Chen, University of Oxford, UK

Appendix A: An Information-Theoretic View of ODDV

As described in Section 4, in ODDV, many visual representations exhibit phenomena of data removal (e.g., node filtering, edge bundling, etc.) or data distortion (e.g., grid maps, metro maps, etc.). In terms of Shannon entropy [Sha48, CT06], these phenomena all feature information loss as illustrated in Figure 9. While it is easy to reason about the demerits of information loss in these phenomena, information theory is so far the only mathematical framework that offers an explanation about the merits of the information loss. Chen and Golan propose to measure the cost-benefit of data analysis and data visualization processes by considering the trade-offs of information loss [CG16]. The qualitative version of the measure is defined as:

$$\frac{\text{Benefit}}{\text{Cost}} = \frac{\text{Alphabet Compression} - \text{Potential Distortion}}{\text{Cost}} \tag{1}$$

where “Alphabet Compression” and “Potential Distortion” are two information-theoretic measures for estimating the positive and negative impacts of information loss caused by a data analysis or visualization process. As the input data and output data of a process may have different forms (e.g., from a time series to a line plot, or from a line plot to perceived visual features), the information spaces of the input and output are referred to as *alphabets* in information theory, while a process is referred to as a *transformation* from one alphabet to another (cf. translation).

When the information space of the output is of less entropy (i.e., less uncertain) than that of the input, there is an information loss, which is a general trend of data analysis and visualization workflows as discussed in detail [CG16]. Hence the impact of this loss is first measured positively as *Alphabet Compression* in Eq. 1. Meanwhile, the negative impact of information loss is measured separately using the term *Potential Distortion*. Chen and Golan proposed to measure the negative impact based on a reverse transformation from output back to input. One may imagine observing data through a visualization image as a reverse transformation from the image back to the original data (cf. a reverse translation in language processing). When one applies the cost-benefit analysis to machine-centric processes (e.g., using statistics and algorithms), one can attest that almost all machine-centric processes also suffer from potential distortion.

This reverse transformation is both **data-dependent** and **user-dependent**, as viewers' knowledge can alleviate the potential distortion. In addition, the potential distortion in a process may or may not impact on the succeeding processes that provide alphabet compression. For example, potential distortion in perceiving an individual data point in a line plot may probabilistically have limited impact on the potential distortion in a succeeding process for determining the trend of a time series. As the cost-benefit analysis can be applied to multiple processes in a workflow, the composite measurement of the potential distortion of a sequence of processes is also **task-dependent** when one views later processes as the tasks of earlier processes in the sequence. The user-dependency and task-dependency is highlighted in Figure 8. A more detailed description of the cost-benefit analysis can be found in a short introduction to the topic at arXiv [Che21].

Consider origin-destination data visualization (ODDV) as an example. Before a viewer observes a visualization image, the viewer is uncertain about the OD dataset D to be displayed. In information theory, all mathematically-valid OD datasets form an alphabet $\mathbb{D}$, which is sometimes referred to as an information space. A valid OD dataset is thus a letter of the alphabet, i.e., $D \in \mathbb{D}$. Every letter in the alphabet is associated with a probability value, $p(D)$, indicating the likelihood that D may appear. In a given context (e.g., rail commuting), many letters in $\mathbb{D}$ become impossible (e.g., about other mode of transport). All possible datasets in this context constitute a sub-alphabet $\mathbb{D}_{\text{ctx}} \subset \mathbb{D}$. In terms of Shannon entropy that measures the amount of uncertainty or information, the entropy of $\mathbb{D}_{\text{ctx}}$ is usually much lower than that $\mathbb{D}$. Knowing the context enables a viewer to think, often unconsciously, using the probability distribution for $\mathbb{D}_{\text{ctx}}$ instead that for $\mathbb{D}$.

When an algorithm is used to manipulate OD datasets in $\mathbb{D}_{\text{ctx}}$, it may further reduce the variations in $\mathbb{D}_{\text{ctx}}$. For instance, as illustrated in Figure 9, node filtering removes the possible variations of those nodes that are deleted if they occur in the data, while edge bundling creates a new alphabet that has fewer letters and thus fewer variations. Grid-mapping and path simplification encode different geometrical variations using the same abstract representation.

In a given context, when a transformation F is applied to all datasets in $\mathbb{D}_{\text{ctx}}$, it results in a new sub-alphabet $\mathbb{D}'_{\text{ctx}}$. If F features operations such as filtering, grouping, or distortion-based abstraction, $\mathbb{D}'_{\text{ctx}}$ will have less entropy than $\mathbb{D}_{\text{ctx}}$. Entropy reduction implies information loss. The usefulness of many visual designs in visualization, such as metro maps and many ODDV designs, evidence that information loss can have an positive impact, while challenging the traditional wisdom that a visual design needs to preserve all information in the data. Sometimes one may argue that a visual design needs to preserve all information useful to a task. While the statement itself captures the task-dependent nature of visualization (but not the user-dependency), it is not ideal as it seems to imply a circular argument: “*a useful visual design shows useful information,*” while neither usefulness can easily be defined.

On the other hand, the cost-benefit analysis proposed by Chen and Golan has offered a mathematical explanation that such visual designs are cost-beneficial. According to the information-theoretic cost-benefit analysis [CG16], such information loss is part of the general trend of entropy reduction in a workflow from a data alphabet to a decision alphabet. Statistics, algorithms, visualization, and interaction in such a workflow all contribute to the entropy reduction (i.e., *Alphabet Compression*). Hence entropy reduction itself is a merit rather than a demerit. Without entropy reduction, there would be no decision.

In addition, entropy reduction at one stage helps reduce the *Cost*

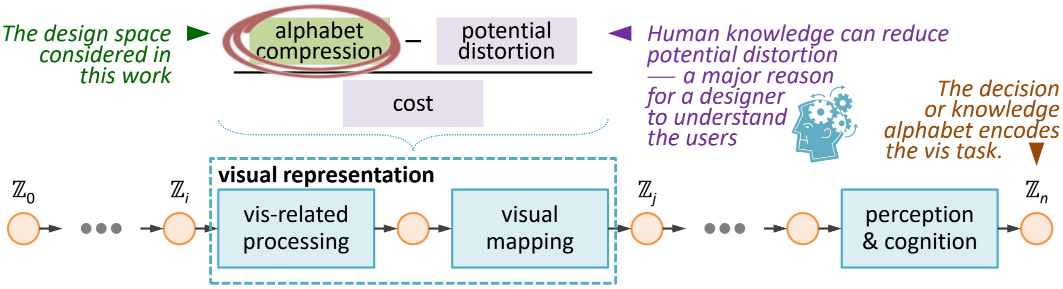

The design space considered in this work
alphabet compression
–
potential distortion
cost
Human knowledge can reduce potential distortion — a major reason for a designer to understand the users
The decision or knowledge alphabet encodes the vis task.
visual representation
$\mathbb{Z}_0$
$\mathbb{Z}_i$
vis-related processing
visual mapping
$\mathbb{Z}_j$
perception & cognition
$\mathbb{Z}_n$

Figure 8: *A design space may categorize different options based on the amount of alphabet compression (i.e., losing information) and ways to achieve it. Too little information loss could increase the cost of the process and slowdown the progress towards the task objective. Too much information loss could increase potential distortion. Users' knowledge can alleviate potential distortion.*

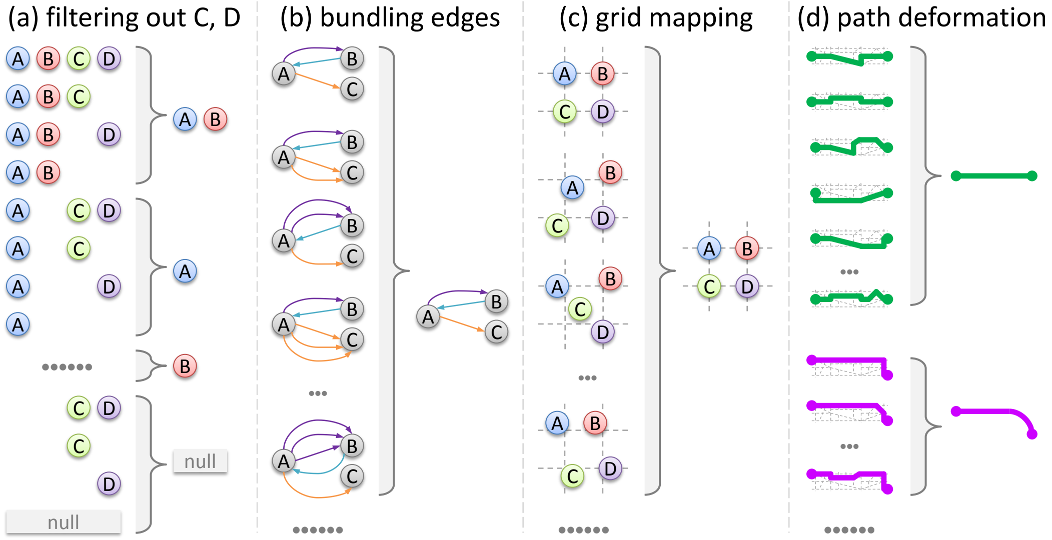


Figure 9: *Four examples of entropy reduction or information loss in ODDV. (a) Whether a dataset may include any of four cities can be defined with an alphabet of 16 letters. When a filtering algorithm removes C and D from any input dataset, it creates a new alphabet with four letters, which has lower entropy. (b) The alphabet for encoding all possible connection patterns (up to k edges) among three nodes contains many letters. Bundling edges with the same source and destination is a many-to-one mapping, which reduces entropy. (c) Grid mapping and path simplification, which are commonly-used design methods in ODDV, are also many-to-one mappings that cause information loss.*

of the stage or the succeeding stages. In Figure 1, for many tasks (e.g., observing if the directional flows between two locations are similar) cost less time or cognitive load with the visual encoding on the right due to the reduction of some informative representations.

Meanwhile, information loss may have a side-effect. When a viewer observes an ODDV image that features filtering, grouping, distortion, or other data transformations that cause information loss, there is a possibility of misinterpretation (i.e., *Potential Distortion*). Using Figure 9(d) as an example, a viewer who has little knowledge about metro maps, may interpret the path between the two stations is straight; a viewer, who understands concept of abstraction but knows little about the geography about that region, may make a random guess that the path can be of an arbitrary shape; or a viewer who lives nearby, may choose a shape that close to the reality. Hence, the misinterpretation is viewer-dependent or user-dependent as we often say in visualization. With the visual encoding on the right of Figure 1, once viewers know that the small doughnut chart at each node summarizes the outgoing flows, they can infer that those nodes without attached lines have little incoming flow. They thus do not suffer much potential distortion that could be caused by not drawing the first half of edges.

In many applications, some types of misinterpretations may not have a negative impact on the succeeding processes, where the transformations would converge to the same decisions regardless the variations of such interpretations. As succeeding processes include tasks, this indicates that visualization is task-dependent. For example, although every edge was drawn fully in top-left image in Figure 1, tracing flow lines are not very effective. This suggests that errors in tracing flow lines may not be an issue with many tasks associated with a generic flow map, providing a rationale for the visual encoding on the right.

Once we appreciate that ODDV should enable entropy reduction and cannot avoid information loss unless the dataset is trivially simple, the question is then about **what** information to lose and **how** to lose information. The principle design criteria are to reduce the potential distortions by maximizing the use of viewers' knowledge, reduce the costs of other human- and machine-processes that handle the data following the information loss, and reduce the negative impact on such processes. In the main body of this paper, we outline a design space categorized based primarily on the notions of **what** and **how**.

Our design space (Section 4) focuses on different ways of alphabet compression as highlighted in Figure 8. In order to organise many ways of entropy reduction or information loss into a design space, we find that the four dimensions discussed in Section 4 can help differentiate ways of entropy reduction and provide the design space with a structure. For Dimensions 1 and 2, it is relatively obvious to consider the transformations Filter and Group as entropy reduction methods. Note that a Filter transformation reduces entropy by removing valid letters in an alphabet, while a Group transformation changes the ordinal alphabet to a new alphabet.

There are also Add and Split transformation in Dimensions 1 and 2. Although entropy reduction is the general trend in the workflow from data to decisions, in many circumstances, a viewer may perform actions to reintroduce some entropy within a particular part of the workflow. In the context of ODDV, a location or path may be added, or a grouped component may be split. Almost all actions for increasing entropy involve interaction. For example, a viewer may sense that the information loss in an overview may cause too much potential distortion, and decide to use zoom-in or tool-tips to bring some details back (i.e., undo filtering); or a viewer may judge that some clustering was not helpful, and exercise some control to split the clusters concerned (i.e., undo grouping). In terms of information theory, the existence of such operations evidences the quantifiable values of human-computer interaction, and demonstrates that interaction enables viewers to self-optimize the amount and pace of information loss during visualization [CE19].

For Dimensions 3 and 4, we find that traditional terms for describing visual encoding of individual nodes and edges do not naturally define categories of entropy reduction. This is likely because the original node list and edge list are the referencing benchmarks for Dimensions 1 and 2 respectively, while it is less common to consider a referencing benchmark for visually encoding a node or an edge. We therefore define *node norm* and *edge norm* as the referencing benchmarks for Dimensions 3 and 4, and then define types of entropy reduction in relation to these benchmarks.

The actions for design space exploration outlined in Section 5.2 are based on a workflow optimisation methodology underpinned by information theory [CE19]. As most filtering and grouping operations are implemented using algorithms, they may have too much alphabet compression to some users, causing symptoms of a high level of potential distortion and/or cognitive load for some tasks. One balancing act is to use interaction as a remedy to to reintroduce the lost information, through interaction itself would introduce extra cost. Similarly, one may use remedies of statistics and visualization to preserve some information (e.g., computing and visual encoding the group size of a super-node or super-edge). Meanwhile, the commonly-adopted wisdom of "knowing the users and tasks" is also supported by the information-theoretic reasoning as shown in Figure 8.

Mathematical Definitions of the Cost-benefit Measure

In the remainder of this appendix, we provide a concise summary of the mathematical definitions related to the cost-benefit measure proposed by Chen and Golan [CG16]. From these definitions, those readers who are knowledgeable about the fundamental concepts in information theory can quickly notice that the cost-benefit measure is composed of two commonly-used information-theoretic measures. For those readers who are new to information theory, these definitions provide a pointer to relevant part of an information theory textbook (e.g., [CT06]). In addition, the original paper by Chen and Golan [CG16] provides the mathematical rationale for the cost-benefit measure, while a recent book chapter by Viola et al. [VCI20] provides the concept of "visual abstraction" with a mathematical explanation based on the cost-benefit measure.

Let $\mathbb{Z} = \{z_1, z_2, \dots, z_n\}$ be an alphabet and z_i be one of its letters. $\mathbb{Z}$ is associated with a probability distribution or probability mass function (PMF) $P(\mathbb{Z}) = \{p_1, p_2, \dots, p_n\}$ such that $p_i = p(z_i) \geq 0$ and $\sum_1^n p_i = 1$. The **Shannon Entropy** of $\mathbb{Z}$ is:

$$\mathcal{H}(\mathbb{Z}) = \mathcal{H}(P) = -\sum_{i=1}^{n} p_i \log_2 p_i \quad \text{(unit: bit)}$$

Here we use base 2 logarithm as the unit of bit is more intuitive in the context of computer science and data science.

An alphabet $\mathbb{Z}$ may have different PMFs in different conditions. Let P and Q be such PMFs. The **Kullback-Leibler divergence** (KL-Divergence), $\mathcal{D}_{KL}(P||Q)$, measures the difference between the two PMFs in bits:

$$\mathcal{D}_{KL}(P||Q) = \sum_{i=1}^{n} p_i \log_2 \frac{p_i}{q_i} \quad \text{(unit: bit)}$$

$\mathcal{D}_{KL}(P||Q)$ is referred as the divergence of P from Q. This is not a metric since $\mathcal{D}_{KL}(P||Q) \equiv \mathcal{D}_{KL}(Q||P)$ cannot be assured.

Consider a transformation $F : \mathbb{Z}_{\text{in}} \to \mathbb{Z}_{\text{out}}$, where $\mathbb{Z}_{\text{in}}$ is the input alphabet of F with a PMF P_{in} and $\mathbb{Z}_{\text{out}}$ is the output alphabet with a PMF P_{out} . The term *Alphabet Compression* in Eq. 1 is the difference between the input and output alphabet, $\mathcal{H}(\mathbb{Z}_{\text{in}}) - \mathcal{H}(\mathbb{Z}_{\text{out}})$.

Consider a reverse transformation F^{-1} that attempts to reconstruct the input from the output. The reconstructed alphabet is expected to have a PMF different from that of the original input alphabet. We denote the reconstructed alphabet as $\mathbb{Z}'_{\text{in}}$ with a PMF P_{in}. Thus the reverse transformation is $F^{-1} : \mathbb{Z}_{\text{out}} \to \mathbb{Z}'_{\text{in}}$. The potential distortion is defined using the KL-divergence as $\mathcal{D}_{KL}(\mathbb{Z}'_{\text{in}}||\mathbb{Z}_{\text{in}})$.

The mathematical definition of the qualitative formula in Eq. 1 is thus:

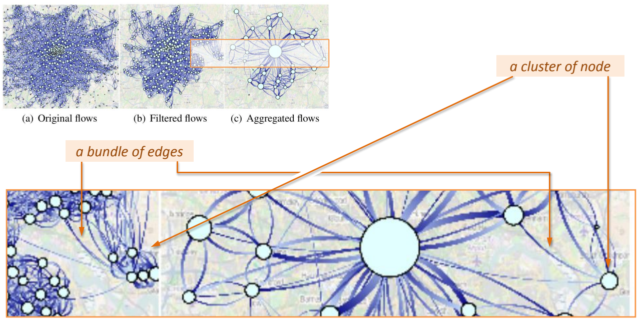
$$\frac{\text{Benefit}}{\text{Cost}} = \frac{\mathcal{H}(\mathbb{Z}_{\text{in}}) - \mathcal{H}(\mathbb{Z}_{\text{out}}) - \mathcal{D}_{KL}(\mathbb{Z}'_{\text{in}}||\mathbb{Z}_{\text{in}})}{\text{Cost}} \tag{2}$$

The fundamental measurement of the Cost is the energy required to perform F and F^{-1}, while it can be approximated by a time or monetary measurement.

Most measurement systems are not ground truth. They are functions that map some reality to some quantitative values, in order to aid the explanation of the reality and the computation of making predictions. The cost-benefit measure in Eq. 2 is one of such functions. While the cost-benefit measure successfully captures tradeoffs in data analysis and visualization workflows, the measured values could shoot up toward infinity easily, hindering the reconstruction of the reality from the measured values. Recently, Chen an Sbert proposed to replace the KL-divergence in Eq. 2 with a bounded divergence measure [CS21], and Chen et al described two empirical studies for collecting practical data and using the data to evaluate several candidate divergence measures [CARSS21]. One of the empirical studies used two London underground maps, one abstract and one geographically-faithful, as the stimuli.

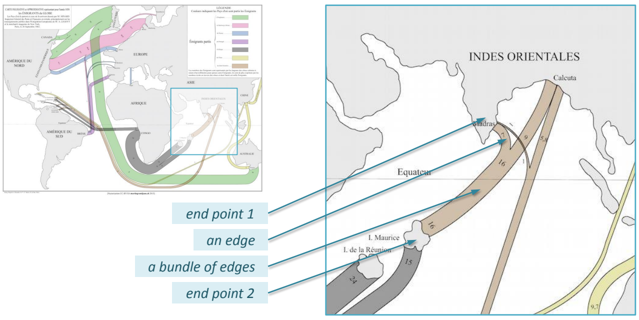
Appendix B: Classification of ODDV in the Literature

The following figures show some examples of ODDV in the literature, and their classification according to the types of transformations in each dimensions.



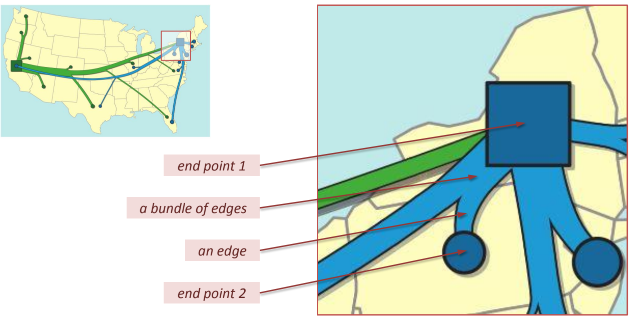
F_1: Group
F_2: Group and filter
F_3: Dimension Enhancement and Attenuation of (x,y)
F_4: Dimension Attenuation w.r.t. ordering,
dimension Attenuation w.r.t. direct path and length

Figure 10: *Example (a): MobilityGraphs [LBR* 16].*



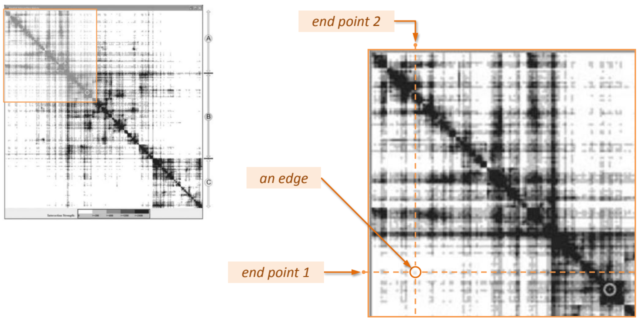
F_1: Group (same coordinates)
F_2: Group (same nodes)
F_3: Geometric Deformation
F_4: Dimension Enhancement,
dimension Attenuation w.r.t. direct path and length

Figure 11: *Example (b): Flow map by Minard (1862) [Rob67].*



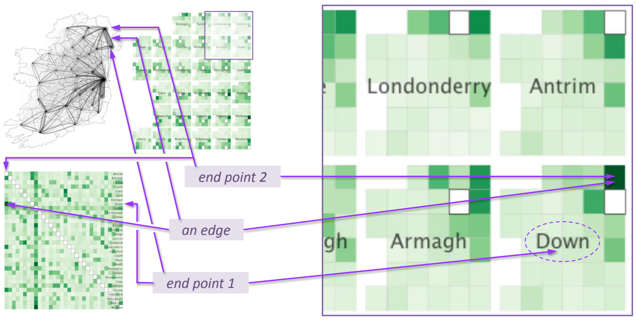
F_1: Group (by admin area)
F_2: Group (same nodes)
F_3: Dimension Enhancement and Attenuation of (x,y)
F_4: Dimension Enhancement

Figure 12: *Example (c): Spiral trees [VBS11].*



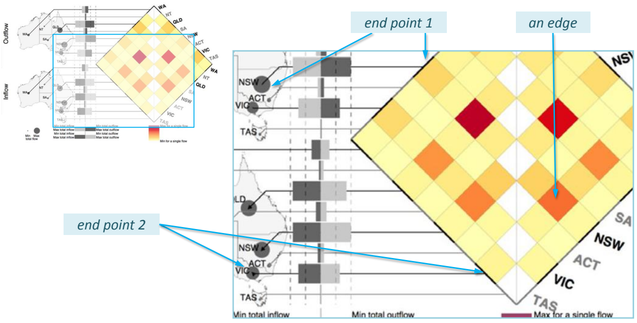
F_1: Group (by admin area)
F_2: Group (same nodes) and add
F_3: Dimension Replacement
F_4: Dimension Reduction

Figure 13: *Example (d): OD matrix [Guo07].*



F_1: Group (by grid cell)
F_2: Group (same nodes) and add
F_3: Geometric Deformation
F_4: Dimension Reduction and Dimension Replacement (right)

Figure 14: *Example (e): OD maps [WDS10, SKD14, cGS* 19]*



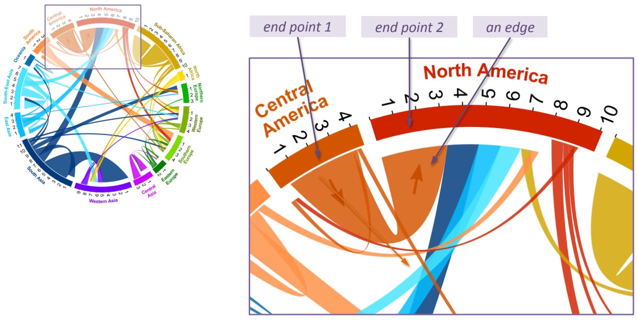
F_1: Group (by admin area)

F_2: Group (same nodes)

F_3: Dimension Enhancement and Attenuation of (x,y) (left)

F_4: Dimension Reduction

Figure 15: *Example (f): MapTrix [YDGM17]*



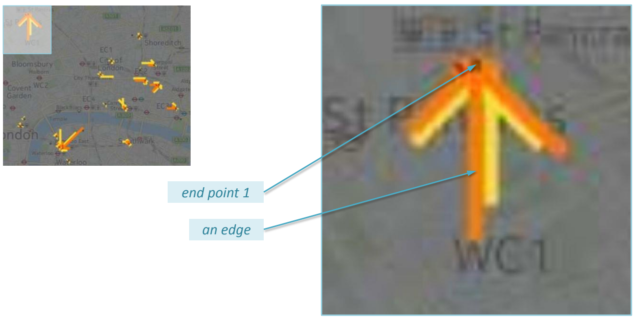
F_1: Group (subcontinent)

F_2: Group (same nodes) and add

F_3: Dimension Replacement

F_4: Dimension Attenuation (w.r.t. ordering) & reduction

Figure 16: *Example (g): Circular plot [AS14]*



F_1: Group (cluster and same nodes)

F_2: Group (angle and length)

F_3: Dimension Reduction (w.r.t. destination node)

F_4: Resolution Reduction (angle)
Dimension Replacement (length)

Figure 17: *Example (h): Flow diagrams [AAFW17]*