AALTO UNIVERSITY
SCHOOL OF SCIENCE
Department of Information and Computer Science
Degree Programme of Bioinformation Technology

Antti Heikkilä

Information Visualisation in a Peer Support Application

Master's thesis

Espoo, August 27, 2012

Supervisor: Timo Honkela, PhD, Adjunct Professor
Instructor: Krista Lagus, D.Sc.(Tech.), Mathias Creutz, D.Sc.(Tech.)

Aalto University
School of Science

AALTO UNIVERSITY
SCHOOL OF SCIENCE
Department of Information and Computer Science
Degree Programme of Bioinformation Technology

ABSTRACT OF
MASTER'S THESIS

Author	Antti Heikkilä	Date	August 27, 2012
		Pages	v + 65

Title of thesis	Information Visualisation in a Peer Support Application
Professorship	Information and Computer Science
Code	T-61
Supervisor	Timo Honkela, PhD, Adjunct Professor
Instructor	Krista Lagus, D.Sc.(Tech.), Mathias Creutz, D.Sc.(Tech.)

Using visualisations to present multidimensional data may help to understand complex relations and to make better decisions. This thesis presents methods for visualising peers based on their similarity. The purpose of the visualisation is to help users of an online peer support service to browse and find relevant peers that are most similar to them.

Four nonlinear dimensionality reduction methods are used to produce visualisations from multidimensional data. The Neighbour Retrieval Visualiser (NeRV), Multidimensional Scaling (MDS), the Self-Organising Map (SOM) and the Generative Topographic Mapping (GTM) are presented and compared quantitatively. The results from the comparison suggest that any one of the four methods could be used in such a peer support service. The methods are then used to visualise data in a hypothetical peer support service called the Stress Map. To further test the methods, the visualisations are subjected to a user study. The visualisation based on the NeRV algorithm performs best, whereas the visualisations made with the SOM and the GTM are judged less appealing.

Keywords

Visualisation, Peer Support, Multidimensional Scaling, Neighbour Retrieval Visualiser, Self-Organising Map, Generative Topographic Mapping

Aalto-yliopisto
Perustieteiden korkeakoulu

AALTO-YLIOPISTO Perustieteiden korkeakoulu Bioinformaatioteknologian tutkinto-ohjelma	DIPLOMITYÖN TIIVISTELMÄ
Tekijä: Antti Heikkilä	Päiväys: 27. elokuuta 2012 Sivumäärä: v + 65
Työn nimi: Informaation visualisointi vertaistukipalvelussa	
Professuuri: Tietojenkäsittelytiede	Koodi: T-61
Työn valvoja: Fil.tri Timo Honkela, dosentti	
Työn ohjaaja: Tekn.tri. Krista Lagus, Tekn.tri Mathias Creutz	

Moniulotteisen datan visualisointi voi auttaa päätöksenteossa, kun se edellyttää monimutkaisten relaatioiden ymmärtämistä. Tässä diplomityössä on esitelty metodeja, joilla voidaan visualisoida ihmisten samankaltaisuutta. Visualisaatioiden tarkoituksena on auttaa käyttäjiä selaamaan ja löytämään itselleen relevantteja vertaisia, jotka ovat mahdollisimman samankaltaisia heidän kanssaan.

Moniulotteinen data visualisoidaan käyttäen neljää epälineaarista dimensionreduktiomenetelmää: Naapurihaun visualisoija (NeRV), moniulotteinen skaalaus (MDS), itseorganisoiva kartta (SOM) ja generatiivinen topografinen kuvaus (GTM). Menetelmien esittelyn jälkeen niitä vertaillaan kvantitatiivisesti. Vertailun tuloksena esitetään, että menetelmät soveltuvat samankaltaisuuden visualisointiin vertaistukipalvelussa. Kuvitteellinen vertaistukipalvelu StressMap esitellään em. menetelmien avulla luotujen visualisaatioiden avulla, jonka jälkeen visualisaatioiden käyttökelpoisuutta testataan käyttäjäkyselyssä. NeRV:iin perustuva visualisaatio pärjää testissä parhaiten, sillä useat käyttäjät vierastavat SOM:illa ja GTM:lla luotuja visualisointeja.

Avainsanat: visualisointi, vertaistuki, vertaistukipalvelu

Preface and Acknowledgements

I wish to express my special thanks to The Department of Information and Computer Science in Aalto University School of Science and to the VirtualCoach research project for providing the funding necessary to complete this thesis. I am particularly grateful to my instructors Krista Lagus and Mathias Creutz, who assisted me in many ways during the process. Warm thanks to Krista for advice and ideas and for the opportunity to work in the project. Many thanks to Mathias, who inspired and encouraged me especially in the beginning of the process. Your regular visits were greatly appreciated.

My grateful thanks to Timo Honkela, who not only supervised the completion of this thesis, but offered the most valuable advice and inspirational discussions. I would also like to offer my great appreciation to the whole VirtualCoach research and partner team, especially Juho Saari for providing me with material regarding peer support, Tuula Styrman, Oili Kettunen and Veera Mustonen for support and inspiration.

I would especially like to thank my colleague and friend Tommi Vatanen for inspirational discussions, instructions and peer support. Thanks to Lassi Haaranen, Jussa Klapuri and Ilari Nieminen for discussions and suggestions on the peer support service. I also wish to acknowledge the valuable feedback I received from Chris, Anna, Cillian, Heikki and Antti, who all reviewed parts of the thesis. Thanks for all the support I've received from my friends in Teekkarispeksi. I would never have finished my studies without the special community you let me be part of.

Finally, I am grateful to my family: Mikko, Marjut and Michele for providing me the foundations and to Elina for bringing grace to my days.

Espoo, August 27, 2012
Antti Heikkilä

Contents

Glossary

Abbreviations

BMI	Body Mass Index
GTM	Generative Topographic Mapping
MDS	Multidimensional Scaling
NeRV	Neighbour Retrieval Visualiser
PCA	Principal Component Analysis
SOM	Self-Organising Map

Chapter 1

Introduction

Imagine the digital stream of data originating from your behaviour, your actions and choices, following you through your daily routines, leaving traces here and there, and never ceasing. Now think what that constant flow of information could tell you, if you could tap into it and reveal the patterns and rhythms of your life. Perhaps you could answer questions like, why did you sleep so well last night, what was it that made you frown in the morning bus? Maybe you could find out when and where you are most happy, what the most significant cause of stress is in your life, and whether you really are satisfied with your job. Think what you could learn if you could further tap into other people's streams and take a look at how they go about doing their things? This sort of hyper awareness might sound like a distant future, but the idea of measuring and analysing ourselves and sharing it with others is becoming more popular along with the development of the Internet.

The Internet, the new ubiquitous medium not only gives access to unlimited information, but offers vast possibilities of connecting with people all around the globe despite any kind of barrier, be it geographical, political or physiological. This immense sense of connection enables people to share their experiences and seek support from one another with unprecedented efficiency. Joys and sorrows can be shared with distant relatives or like-minded peers in real time despite the difference in time or location. An individual with a rare medical condition can easily find people in a similar situation, even if there was no one else in the same country with the same ailment. Online support groups also cover a range of issues beyond medical conditions such as parenting, bereavement or loneliness.

None of this would be possible without the computer. The amount of data, its dimensionality and the potential ways of interacting with data, have all increased due to the use of sophisticated computer systems. At the same time the information and the important relations inherent in the data have become more complex. Human beings are poor at processing multidimensional data and it is challenging even for an expert to see all the relevant relations in complex data. However, many decisions could be improved if all of the available data could be taken into account. Using just a subset of the available variables might lead to suboptimal decisions. One way to ease the cognitive load of processing multidimensional data is to use visualisations. The idea of presenting multidimensional data visually is to externalise information processing from the symbolic and logical inferences to perceptual ones. Externalisation frees the user's cognition by outsourcing the information processing to other means. [5, p. 2–8]

Health, wellbeing and maintaining the ability to work are key challenges for the society and the national economy. The development of preventative services and tools that support the empowerment of individuals from their subjective starting points is the objective of the TEKES funded VirtualCoach project. VirtualCoach aims at introducing statistical data analysis and machine learning in wellbeing research to develop new kinds of tools and services. The project is carried out as a collaboration between the Aalto University School of Science, Aalto University School of Economics, the National Consumer Research Centre, University of Eastern Finland as well as a number of field experts. This thesis forms a part of the VirtualCoach project. [3]

The aim of the thesis is to study the use of computational visualisation methods in visualising the peer relationships of a set of people in a hypothetical peer support service called the Stress Map. The specific research questions are: 1. what does it mean to visualise peers, 2. which kinds of methods could be used to visualise and 3. what should be taken into account when visualising peers with these methods? The concept of peer support and the pursuit of finding relevant peers are discussed as well. The main contribution of the thesis is the design and evaluation of visualisations for a suggested peer support service called the Stress Map. The visualisations are subjected to evaluation in a small-scale user survey conducted with an online web form. The implemented visualisation methods are also compared quantitatively and suggestions are made.

Developing the topic of the thesis has been a multistage process. Initially, there was a set of wellbeing related data sets and an ambition to make a connection between the machine learning and statistical analysis community and the wellbeing and peer support community. The starting point was a stress related questionnaire data set gathered at the 2011 Wellbeing Innovation Camp (http://www.cis.hut.fi/wicamp/) as a part of the VirtualCoach project. The objective was to find possible stress profiles in the data with the help of experts from the VirtualCoach group. A workshop was prepared and held at the "Ski-camp" at the Sport Institute of Finland in Vierumäki in March 2012. At the workshop a tentative analysis of the data was presented along with some visualisations. The workshop participants were experts from various fields e.g. economics, coaching, peer support and entrepreneurship. The experts familiarised themselves with the data and the visualisation methods and discussed the data, the possible profiles and the idea of visualising peer relationships. Based on the results of the workshop, the direction of the thesis was changed and a more concrete approach was adopted. The focus was turned from the data to the visualisation methods and a hypothetical peer support service was designed. The service was also partly inspired by the Questionnaire prototype designed in the VirtualCoach project [23]. However, the aim of establishing a connection between two fields of science remained.

In peer support, feelings of loneliness, rejection, discrimination and frustration are battled with support, companionship, empathy, sharing and assistance offered by individuals who share similar experiences. It is not well known how and why peer support works, but several underlying psychosocial processes are suggested to explain at least some of the benefits, e.g., social support, experiential knowledge, social learning theory, social comparisons and the helper-therapy principle [40]. Online support groups are numerous and serve a diverse field of subjects. However, the difficulty is now in finding a relevant group and establishing meaningful contacts inside the groups. Statistical analysis methods can help in the task.

Dimensionality reduction methods are powerful tools for treating multidimensional data and producing visualisations. Linear methods are well understood and widely used, but they are not always useful for analysing complex data with sophisticated relationships. Numerous nonlinear dimensionality reduction methods exists which all have their advantages and disadvantages. Four such methods were studied in this thesis. The Multidimensional Scaling (MDS) [42] was chosen because it is an old and widely used method. MDS transforms proximity measures into distances on a lower dimensional space. A recently developed method called the Neighbour Retrieval Visualiser (NeRV) [46] was chosen as a state of the art method to be compared with the MDS. NeRV aims at producing visualisations that help to retrieve the neighbourhood structures of the original data as well as possible. The next two methods were chosen from the field of topographic mapping. The Self-organising map (SOM) [25] is a successful method of nonlinear dimensionality reduction that orders multidimensional data on a 2-dimensional map topographically, so that similar data points lie close to each other. The Generative Topographic Mapping (GTM) [7] is a similar method, but derived from more probabilistic principles. In this thesis, these methods are compared quantitatively and then implemented to produce visualisations.

The visualisations are used in the design of a mock-up prototype of a peer support service called the Stress Map. The prototype serves to express the idea of finding peers based on visualisation. To learn more of the methods, the visualisations were tested on an audience that was not part of the VirtualCoach project. A small-scale survey was conducted using a web form with static example visualisations of the service. The results of the survey suggest some guidelines to take into account when using these kinds of visualisations.

The thesis is organised as follows. Peer support and similarity are discussed in Chapter 2. In Chapter 3 the visualisation methods are presented and compared. The implemented visualisations along with the Stress Map service are presented in Chapter 4. Readers not interested in the technical details of the visualisation methods are recommended to skip Chapters 2 3 and continue reading in Chapter 4. The user survey is found in Chapter 5 and finally Chapter 6 concludes the thesis.

Chapter 2

Peer Support and Similarity

When feeling depressed, hopeless or facing the illness of a loved one, it is natural to seek help from others who have been through similar experiences. In situations like coping with stress, breaking up with a partner or even the birth of a child, peer support may offer empathy, companionship and assistance. Why does it help to share your problems with others and, what is it exactly that makes you regard someone as similar to you?

This thesis relies on research findings (see, e.g., [9, 6, 38, 41]) that show peer support is beneficial for human beings and that having respectful relationships with similar people or people in a similar situation can increase personal wellbeing. It is also assumed that this similarity of people can somehow be formalised and measured and used in finding people that match to each other. In this section, these assumptions are examined and justifications are presented.

2.1 Peer support

Peer support is a system of giving and receiving help founded on key principles of respect, shared responsibility, and mutual agreement of what is helpful [35]. Gartner and Riessman [13] define peer support from a more mental health centered perspective: "peer support is social emotional support, frequently coupled with instrumental support, that is mutually offered or provided by persons having a mental health condition to others sharing similar mental health conditions to bring about a desired social or personal change." According to Mead et al. [35], what is normal is defined by the cultural mainstream enforcing a patient identity to people labelled with mental illness or other disabilities. In peer support, people are able to find affiliation with others they feel are similar to them. Peer support at best can offer a culture of health and ability as opposed to a culture of illness and disability. The effects of peer support have been studied and the benefits to the receivers and the providers of peer support are reported in numerous studies [40, see p.395–396].

Peer support is closely associated with the more general theory of social support studied in psychology and sociology. Social support is proven to reduce morbidity and mortality, lessen exposure to psychosocial stress and perhaps other health hazards and buffer the impact of stress on health [19]. However, researchers do not agree on what exactly it is about social support and social relationships that affects health and how it works.

While the concept of social support is found to be multidimensional and problematic, Vangelisti [45] separates three important aspects: *received support*, *perceived support* and *enacted support*. Received support is what individuals get from their social environment, while perceived support refers to the type or amount of support that individuals believe is available to them. Enacted support refers to the verbal and nonverbal behaviour that individuals engage in when trying to provide someone with help.

In addition to social support, Solomon [40] recognises a variety of psychosocial processes that are believed to explain the benefits of peer support. Experiential knowledge is specialised information and perspectives that people obtain from living through similar experiences. Social learning theory states peers make more credible role models for others with similar experiences and therefore interactions with peers are more likely to result in positive behaviour change. Social comparison theory distinguishes two types of social comparison. Upward comparison is considered to provide other peers with an incentive to develop their skills and to offer hope, while downward comparison can put in perspective how bad things could be and how much better off an individual is compared to his or her peers. Finally the helper-therapy principle states that individuals benefit from helping others.[40]

The surge of online support groups has been enabled by the development of the Internet in the last decades. This new kind of peer support has gained wide popularity and groups have been formed around diverse subjects. Most online support groups are set up by peers and not health care professionals. While the number and diversity of online peer support increases, it becomes increasingly difficult to find the right forum among the plethora of possibilities. One can perhaps find a group, but finding a good one with actual high quality conversation is much more difficult.

Even in a particular support group it might be difficult to find relevant stories or experiences. How does one find the relevant peers who have had similar experiences? Lagus and Saari [31] propose a framework in which statistical analysis is used for identifying potential peer groups, in order to suggest peers. The next section discusses the analysis of similarity between individuals.

2.2 Similarity and closeness

To claim Pekka is a peer of Jaakko implies some form of likeness between the two. They might work in the same company or field, have similar physical appearance, belong to the same age group or share common interests or political views. Perhaps they suffer from the same ailment or maybe they just went to the same high-school or live on the same street. In general, they have something in common. Whether they would like to engage in any kind of interaction probably depends at least partly on how similar they consider to be to each other. All of the above mentioned factors could be taken into account when determining how similar Pekka and Jaakko are. Performing this kind of matching of features and judging the amount of likeness between two objects is called measuring similarity.

Similarity is an important concept in many fields of science, although different aspects of it are studied in different fields. In information retrieval, similarity appears as a mathematical construct, whereas in the social sciences it is approached from a more philo-

sophical point of view. Medin et al. [36] argue that similarity seems to have no meaning unless it is specified carefully. Any two things share an arbitrary number of predicates and differ from each other in an arbitrary number of ways. Thus the predicates need to be constrained to be able to judge the similarity between any two objects. Defining similarity is to study these constraints according to Medin et al. [36].

One of the oldest approaches to similarity is the geometric model or psychological distance assumption [10]. The model describes a percept as a mental representation of an object, a concept, an ideal or any other mental entity that can be quantified. Percepts are assumed to vary on a set of features that can be understood as psychological distances and can have numerical values. Thus, a percept is represented by a point in a multidimensional psychological feature space. Similarity between percepts can then be determined by calculating a distance between the points in the mental space. Objects that are close in the space are assumed to be more similar than objects that are located far apart [39]. A multitude of different distance functions exist from which the Euclidean distance and the city-block distance are the most common. Because similarity is expressed as a function of distance, it necessarily has to obey the distance axioms like symmetry $d(A,B) = d(B,A)$ and triangle inequality $d(A,B) + d(B,C) \geq d(A,C)$ [10].

Unfortunately it is easy to come up with counterexamples that question the validity of these axioms. According to Tversky [44], North Korea is usually considered more similar to China than China is to North Korea. Also an ellipse is more similar to a circle than a circle is to an ellipse. Abandoning the metric and dimensional interpretation of similarity, Tversky introduces a feature based similarity measure he calls the *contrast model*. The idea is that similarity is measured through a process of matching two sets of features. Similarity then depends not only on the features the two concepts have in common, but also on the features that are unique to them. The asymmetry of similarity can be achieved by weighting the combination appropriately.

Other approaches to similarity include the structural alignment-based models and the transformational models. In general, similarity is difficult to measure because the judgements are influenced by context, perspective, choice alternatives and expertise [36]. Choosing which features are used when two concepts are compared is an important question when trying to make similarity judgements.[14]

This thesis studies the similarity between people as peers, a particularly complex and abstract type of similarity. In order to visualise the similarity between people, the similarity needs to be computed making the distance-based model a natural choice. Still, the question remains, which of the psychological dimensions should be studied and what the set of features is that should be considered when determining similarity between people.

Selecting features to preprocess the data

The task of choosing a suitable set of features in a representation is called feature extraction in the machine learning literature. This preprocessing step is such an integral phase of the visualisation process that it is often taken for granted [29]. However, as Kaski [20, p. 34] points out, the visualisation methods merely illustrate some structures that are ultimately determined by the features chosen to represent the data items. The task of tailoring the desired features to reflect the requirements of a given application requires usually considerable expertise not only in the application area but in the data analysis

methodology as well.

In order to illustrate the process further, let us formulate a hypothetical feature vector $\mathbf{x}$ that represents different dimensions of an individual:

$$\mathbf{x}^{\mathrm{T}} = (\underbrace{A,B,}_{\text{fitness}}\underbrace{C,D,E,}_{\text{health}}\underbrace{F,G,}_{\text{relationships}}\underbrace{H,I,}_{\text{work}}\underbrace{J,K,L,}_{\text{interests}}\underbrace{M,N,O}_{\text{values}})$$

and let's assume these variables can be somehow measured, e.g., directly with physiological tests or indirectly with the use of questionnaires or other types of measures. Now the similarity s between two persons $\mathbf{x}$ and $\mathbf{y}$ may be calculated with a suitable function:

$$s(\mathbf{x},\mathbf{y}) = f(\mathbf{x},\mathbf{y}), \tag{2.1}$$

where the similarity s can be a scalar or vector valued. In the distance-based similarity models, this function f is usually the Euclidean distance or the city-block distance:

$$\delta^2_{xy} = \left(\sum_{i=1}^{m}(x_i - y_i)^2\right)^{1/2} \qquad \text{Euclidean distance} \tag{2.2}$$

$$\delta^1_{xy} = \sum_{i=1}^{m}|x_i - y_i| \qquad \text{city-block distance,} \tag{2.3}$$

where m is the number of variables and x_i is the observed value of the attribute i of object x.

To have a more flexible measure of similarity the function f can be chosen so that it depends on the context. Even asymmetric functions may be considered depending on the application. $f(x,y)$ doesn't necessarily have to equal $f(y,x)$, although it might be difficult to handle this kind of asymmetric space. One way of obtaining similarities is to calculate correlations between the attributes and the objects. However, if the attributes are standardised the resulting correlation coefficients are related to the Euclidean distance by a monotonic function [8, p. 130].

Even if a simple function like (2.2) is used, the question which set of features is used in the calculation is important. For example, a car and a motorcycle both have wheels, thus the attribute of having wheels has no meaning when comparing cars and motorcycles. However, in a wider perspective, if boats and flying machines are taken into consideration, then the attribute of having wheels does indeed convey useful information. Later in this thesis the similarity between people as peers is considered in the context of stress and relaxation. In that case it is unimportant if the people have different hair colours or if they differ a lot in height, whereas aspects like work, sleep, nutrition, health and family relationships are probably a lot more informative.

In this thesis, the similarity between multiple persons is compared and visualised. So instead of just comparing two vectors $\mathbf{x}_1$ and $\mathbf{x}_2$, whole sets of vectors $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_1, \ldots, \mathbf{x}_n)$ are analysed. Studying the structure of the whole data set imposes other conditions on the similarity measure. The similarity between all pairs of objects can be calculated using a suitable function f resulting in a similarity matrix:

$$\mathbf{S}_{i,j} = \begin{pmatrix} 1 & s_{1,2} & \cdots & s_{1,n} \\ s_{1,2} & 1 & \cdots & s_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{1,n} & s_{2,n} & \cdots & 1 \end{pmatrix}$$

Here the similarity of an object with itself is assumed to be scaled to 1 and $s_{i,j} = s_{j,i}$ resulting in a symmetric matrix.

Different sets of features will produce different results. Some features might emphasise the dissimilarity of objects more than the similarity. Others concentrate more on the similarity or shared features, and disregard small differences. If the similarities are visualised as will be seen in Chapter 3, these choices have an effect on the global structure of the data set. Some objects are pushed further away because of their differences and some are grouped together because they are more similar. In this thesis, it is argued that the local relationships are more important than the global structure when visualising individuals for peer support purposes.

Chapter 3

Visualisation Methods

The past few decades have indicated a tremendous change in the way we produce, analyse and use data. The use of computers has increased the amount of data, its dimensionality and the ways of interacting with the data. Yet the new technology has brought new challenges as the significant relations have become more complex. Inspecting multidimensional data represented by tables is difficult even for an expert. Humans have limited cognitive capabilities, hence they are not able to process all the available information, which leads to suboptimal strategies. Extracting relevant information and simplifying data is an important issue for businesses and institutions working with large amounts of data [12].

One way to make sense of multidimensional data is to visualise it. Computers are good with numbers, humans are good with visual patterns. By transforming the raw numerical data into visual representations, the best of both approaches are combined. More data can be taken into account when making decisions, if the important relations in the data are exposed in visual form. Visualisations can overcome mental limitations through a process called externalisation. The idea of externalisation is that perceptual inferences are used instead of logical and symbolic ones. In other words, finding, e.g., an outlier in a graph is easier than in a fact table filled with numbers.

Information visualisation as a practice can be traced back to the earliest geographic maps made in ancient times. Numerical data was placed on such a map for the first time in the 17th century and the more abstract idea of mapping arbitrary data on a graph with no relation to the physical world did not show up before the early 19th century [43]. Today, information visualisation is getting more and more attention from the scientific community as well as from a larger crowd. For example, World Design Capital Helsinki [49] chose information visualisation as one of their key topics and Mehmood et al. [37] used visualisations in the field of wellbeing to illustrate the connections between health and nutrition.

However, information visualisation is never a completely objective process. Thus, each visualisation must be carefully designed to achieve a representation that serves its purpose as well as possible. Multidimensional data cannot be directly visualised and the result is always a compromise of some kind. Choices must be made regarding the amount of detail, emphasis and overall clarity.

The multiple dimensions of the data need to be reduced to two or three in order to visualise the data on a flat surface, hence dimensionality reduction is a necessary pre-

processing step of visualisation. In this thesis, the words dimensionality reduction and visualisation are used interchangeably. Dimensionality reduction relies on the assumption that the data resides on a lower-dimensional manifold or surface embedded in the higher-dimensional space.

The choice of a relevant visualisation method is at the core of this thesis. In this chapter, a set of visualisation methods is presented and compared. The methods can be divided into three categories: linear methods, nonlinear projection methods and topographic map methods. The linear methods and nonlinear projection methods map the data continuously on a two-dimensional plane, while the topographic map methods represent data by matching it with discrete reference vectors on a two-dimensional map. All this will become clear in the following sections.

3.1 Linear Methods

Dimensionality reduction is usually performed by mapping or projecting the original data into a lower-dimensional subspace. This projection can be linear or nonlinear. In a linear projection, the location of a data point can be represented as a linear combination of the original coordinates. This restricts the lower-dimensional manifold to be a linear subspace. In a nonlinear projection, the possible surfaces can have more complex structure. However, a direct mapping of the projected dimensions into the original space may not be possible, thus making the interpretation of the results somewhat more difficult.

A widely adopted linear dimensionality reduction method is the principal component analysis (PCA) [18], which seeks orthogonal projections of the original data so that the maximum amount of information is preserved. This optimal transformation can be found by calculating the eigenvectors of the data covariance matrix $\mathbf{S} = \frac{1}{N}\sum(\mathbf{x}_n - \mu)(\mathbf{x}_n - \mu)^{\mathrm{T}}$. The data is then projected to a k-dimensional subspace using the eigenvectors corresponding to the k largest eigenvalues $\lambda_1, \ldots, \lambda_k$. The transformation is given by:

$$\mathbf{z} = \mathbf{W}^{\mathrm{T}}(\mathbf{x} - \mu), \tag{3.1}$$

where $\mathbf{z}$ is the transformed data vector, μ is the mean of the data and $\mathbf{W}$ is a matrix containing the k chosen eigenvectors from the data covariance matrix. [18]

Figure 3.1 illustrates the use of PCA schematically. Some artificial data is sampled from a bivariate Gaussian distribution. The dashed red line represents the covariance of the distribution. PCA then finds the orthogonal projection that maximises the variance in the data. The two green vectors represent the resulting principal components of the data. If a one-dimensional representation is needed, projecting the data onto the first principal component would preserve most of the variance.

Linear methods like PCA or factor analysis are easy to implement and often it is straightforward to interpret the resulting locations of the data points. However, when visualising complex phenomena like the similarities between human individuals the assumption of an underlying linear manifold may be too simplistic. Also it may be the case that no global features are discernible, and thus linear methods may fail to find any structure in the data. In such cases, it is more important to focus on the local relationships of the data and put emphasis on the neighbourhood structure of the objects. This can be achieved by nonlinear methods, that are discussed next.

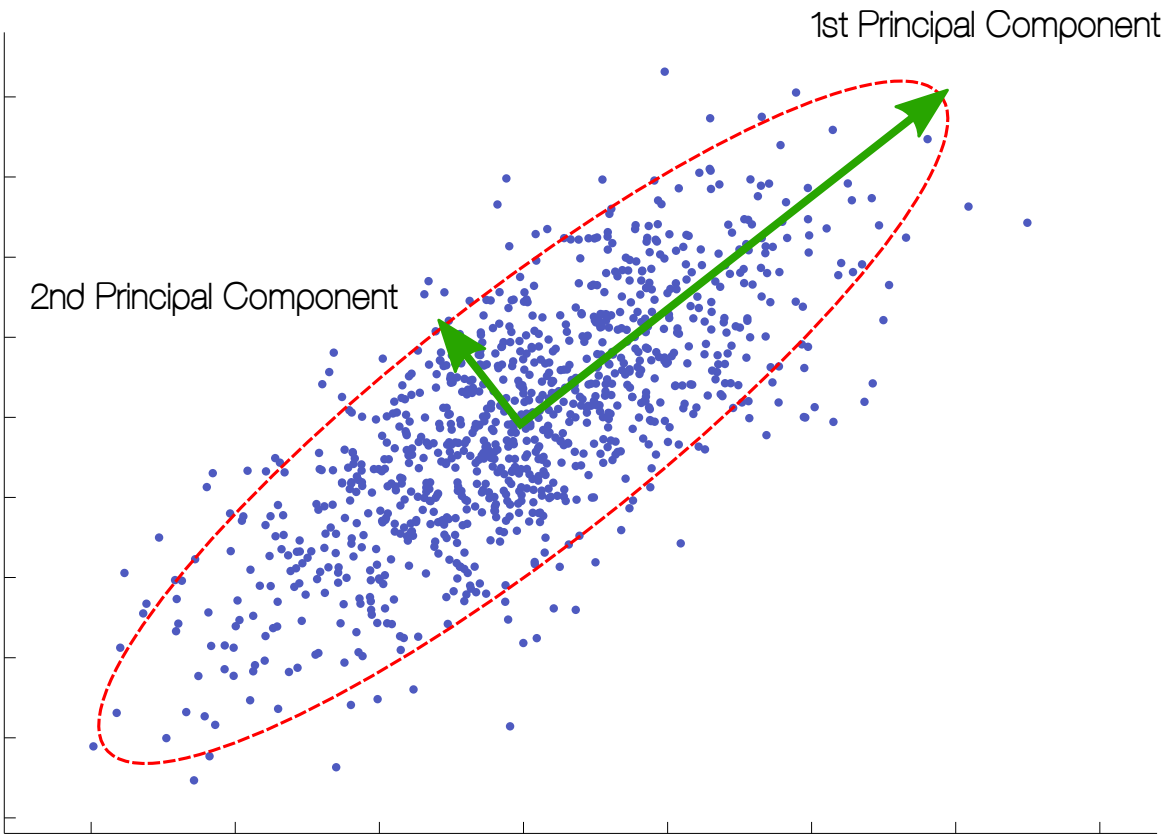


Figure 3.1: A schematic illustration of PCA on a 2-dimensional normally distributed data.

3.2 Nonlinear Projection Methods

Two nonlinear visualisation methods are presented and compared in this section. Multidimensional scaling (MDS) [42] is a traditional method widely used in psychometric research, whereas Neighbour Retrieval Visualiser (NeRV) [46] is a recent state of the art visualisation algorithm from the field of data analysis and machine learning.

3.2.1 Multidimensional Scaling

Multidimensional scaling refers to a set of methods usually used in visualising proximity data. The idea of MDS is to represent observed dissimilarities as distances in a lower dimensional space. The observed objects are mapped as points in the MDS space so that their distances match the original dissimilarities as well as possible. This is useful if only the dissimilarity of the objects is known or if the objects are of high dimension. MDS is often used as a data exploration tool to gather insight into new data.[8, p.3–6]

The Classical MDS as proposed by Torgerson [42] assumes that the original dissimilarities are Euclidean distances resulting in a linear projection of the objects. For non-Euclidean data, e.g., correlation coefficients, interval or ordinal data, a more general approach is needed. There are both metric and nonmetric MDS methods for handling this kind of data.

Extending the idea presented by Shepard [39], Kruskal [27, 28] first formalised MDS by introducing a goodness of fit measure called stress. Following the notation in [15], let

n be the number of objects under study and the observed dissimilarity between objects i and j be given by δ_{ij}. Let $\mathbf{X}$ be the $n \times d$ matrix of coordinates where d is the desired dimensionality chosen by the user. Rows of $\mathbf{X}$ are then the new transformed coordinates of the objects. The Euclidean distance $d_{ij}(\mathbf{X})$ between rows i and j of $\mathbf{X}$ is defined as

$$d_{ij}(\mathbf{X}) = \left(\sum_{s=1}^{d} (x_{is} - x_{js})^2 \right)^{1/2}. \tag{3.2}$$

The idea of MDS stated once more is to find the $\mathbf{X}$ such that $d_{ij}(\mathbf{X})$ matches δ_{ij} as well as possible. This goal is usually sought through the use of a cost function called stress. Kruskal [27] introduced raw-Stress given by

$$\sigma^2(\mathbf{X}) = \sum_{i=2}^{n} \sum_{j=1}^{i-1} (\delta_{ij} - d_{ij}(\mathbf{X}))^2. \tag{3.3}$$

Because the dissimilarities δ_{ij} and the distances $d_{ij}(\mathbf{X})$ are symmetric, the summation only involves the pairs where $i > j$. However, the raw-Stress is not invariant under stretching or shrinking of the configuration. A better measure is found by normalising and taking the square root

$$\sigma_1(\mathbf{X}) = \sqrt{\frac{\sigma^2(\mathbf{X})}{\sum d_{ij}^2(\mathbf{X})}} = \sqrt{\frac{\sum_{i>j} (\delta_{ij} - d_{ij}(\mathbf{X}))^2}{\sum_{i>j} d_{ij}^2(\mathbf{X})}} \tag{3.4}$$

resulting in the second measure called Stress-1 as proposed by Kruskal [27].

Before using MDS, the dissimilarities must be obtained somehow from the data. None of the three datasets examined in this thesis contain direct dissimilarities and thus a mechanism for obtaining these must be chosen. Each of the datasets consist of objects from which several attributes have been measured resulting in an array of attribute profiles. For example the stress survey data, introduced more thoroughly in 3.2.3, is a collection of answers to questions on a Likert-scale ranging from 1 to 5. Choosing the method of obtaining dissimilarities is illustrated using the stress survey data as an example. A portion of the data is gathered in Table 3.1.

Table 3.1: A portion of the stress survey data. The rows of the table correspond to respondents (objects) and the columns to questions (variables).

1	1	1	2	1	2	2	2	3	2	2	3	3	1	1	1	2
1	1	2	1	2	3	4	2	4	4	4	1	2	1	1	1	2
2	4	2	1	1	1	1	1	4	1	4	5	2	3	2	1	2
2	3	2	1	2	3	4	4	5	4	2	2	2	3	2	1	2
1	1	1	1	2	3	2	1	1	1	1	1	1	1	1	1	2

There are several ways to calculate dissimilarities among the respondents. One possible way is to calculate correlations between the objects according to their answers to the

questions. Another way is to calculate distances directly from the attribute vectors. For example, the city-block distance may be used to obtain an estimate of the dissimilarity

$$\delta_{ij}^1 = \sum_{a=1}^{m} |x_{ia} - x_{ja}|, \tag{3.5}$$

where m is the number of variables and x_{ia} is the observed value of the attribute a of object i. However, for the stress survey the Euclidean distance

$$\delta_{ij}^2 = \left(\sum_{a=1}^{m} (x_{ia} - x_{ja})^2 \right)^{1/2}, \tag{3.6}$$

was used because it was assumed that all the variables were equally distributed. A collection of popular distance measures can be found, e.g., in [8, p.122]. Obtaining the dissimilarities from mutual correlations emphasises the trends in the answers of the respondents ignoring the magnitude. On the contrary, calculating the Euclidean distance over the attribute vectors takes the magnitude of the answers into account. This is a service design question. In the stress survey case, it may be argued that taking the magnitude into account is more important. The respondents that have reported high overall values should probably be located close to each other even though their answers would differ a little. On the other hand, respondents who report relatively low values on similar questions should be kept apart from the respondents with high reported overall values. However, this is completely application specific and subject to judgement on the part of the researcher.

In many cases, the dissimilarities δ_{ij} are not known or they are not in an easily comparable numerical form. For example, the dissimilarities might be unknown but their order is observed. In such a case, the dissimilarities must be approximated by some numerical values often called disparities or d-hats. In a nonmetric case like this, MDS tries to simultaneously find the coordinates $\mathbf{X}$ and the approximated disparities. Modifying Stress-1 yields

$$\sigma_1(\mathbf{X}) = \sqrt{\frac{\sum_{i>j} (\hat{d}_{ij} - d_{ij}(\mathbf{X}))^2}{\sum_{i>j} d_{ij}^2(\mathbf{X})}}. \tag{3.7}$$

Obtaining these disparities from the original dissimilarities is called transforming the data. If there exists a continuous function that maps the dissimilarities into disparities the term *metric* MDS is used. If only the rank-order of the dissimilarities is known, then the disparities need to preserve this order and the transformation must be a monotone function and the term *nonmetric* MDS is used.[8, chap.9]

More information about the fit made by MDS can be obtained by examining a scatter plot called the Shepard diagram Borg and Groenen [8, p.42–44]. An example of a Shepard diagram plotted for the Stress survey data can be seen in Figure 3.2. The Shepard diagram presents the fitted MDS distances $d_{ij}(\mathbf{X})$ against the original dissimilarities δ_{ij} as blue open circles. The approximated disparities $\hat{d}_{ij}$ are also plotted against the dissimilarities shown as red filled circles. The vertical distance between each $(d_{ij}(\mathbf{X}), \delta_{ij})$ (blue open

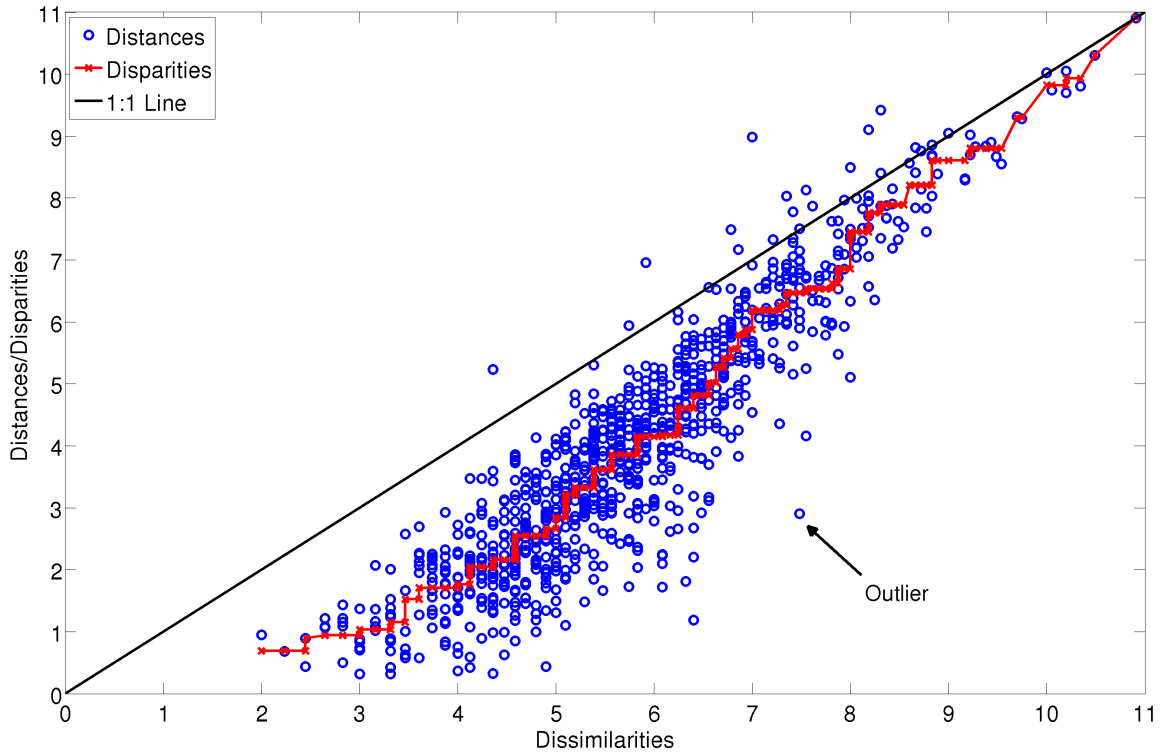


Figure 3.2: A Shepard diagram for the MDS fit of the Stress survey data

circle) and $(\hat{d}_{ij}, \delta_{ij})$ (red filled circle) represents the reconstruction error for a particular value of dissimilarity.[8, p.42–44]

The Shepard diagram in Figure 3.2 shows that there is quite a lot of scattering around the red curve that represents the obtained disparities. There are a few outliers showing significant distortion in the reconstruction. For example the point $(7.5, 2.9)$ represents a significant mismatch between the original dissimilarity and the estimated MDS distance. In other words, there are two objects which are originally estimated to be far away, but are placed close to each other in the MDS space. The whole curve bends below the 1:1 line meaning that the fit tends to underestimate small and moderate dissimilarities. The relationship is not exactly linear either, but that might be acceptable as the plot is aiming at showing the "neighbours" of objects, thus it should be sufficient to preserve the non-metric criteria of rank order. This is further discussed in Chapter 4 where the rationale for this kind of visualisation is presented.

To conclude, MDS refers to a set of methods that represent various kinds of dissimilarities as distances in a lower dimensional space. Usually MDS is used for visualisation purposes where the dimensionality of the resulting space is two or three. Fitting the MDS is usually done by minimising a cost function called Stress. In some cases, the dissimilarities have to be transformed to obtain real valued disparities. In metric MDS, the transformation is a continuous function. If only the rank order of the data is to be preserved, then a nonmetric variations of MDS can be used. The performance of a specific MDS implementation can be assessed by examining a Shepard plot that relates the estimated distances and disparities against the original dissimilarities.

3.2.2 Neighbour Retrieval Visualiser

Visualising multidimensional data on a flat surface always includes a tradeoff of some kind. Take, for example, the age old cartographer's problem of drawing the surface of the Earth on a map. The basic question is, how to present the surface of a three-dimensional sphere on a two-dimensional plane without distorting the image. The result is always a compromise; some parts of the image are stretched and some contracted. A similar situation is visualised in Figure 3.3. Consider data points located on the surface of a three-dimensional sphere. Figure 3.3a presents the original data and below two possible ways of projecting the points on a 2-D plane. In Figure 3.3b, the sphere has been squashed flat. The result represents a linear projection and highlights the need for a more complex treatment. A different approach is presented in 3.3c where the surface has been cut open. The dataset is from the dredviz website [1] as presented in Venna et al. [47].

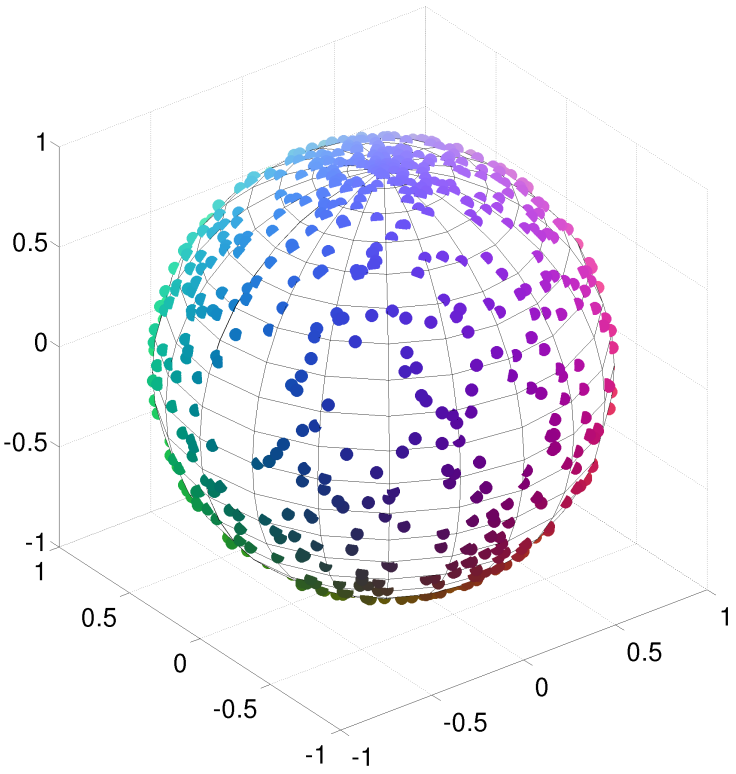
Which one of these two approaches is more correct? It turns out that the answer depends on the data and the purpose of the visualisation. For the purpose of visualising similarity relationships in two dimensions, the task was formulated by Venna and Kaski [46] as an information retrieval problem. They suggest that precision and recall should be used in measuring how well a given visualisation retains the original neighbourhood structure of the data.

The visualisation might miss some neighbours and present some false neighbours as illustrated in Figure 3.4. The shaded rectangle on the left represents the original m-dimensional input space. The output space on the right represents the visualisation in two dimensions. The Figure concentrates around a single sample i, whose location in the input space is given by x_i and in the output space y_i. P_i is the neighbourhood of i in the input space depicted by a circle and Q_i is the corresponding neighbourhood in the output space. Neighbourhood can be a fixed number of nearest neighbours or it can be defined as the set of nearest points within a fixed radius. The shaded circles represent actual neighbours of sample i and stars some other samples. In the process of visualisation, some of the actual neighbours might get lost resulting in misses and reduced recall. Alternatively, some samples might end up close to the sample resulting in false neighbours and reduced precision.[47]

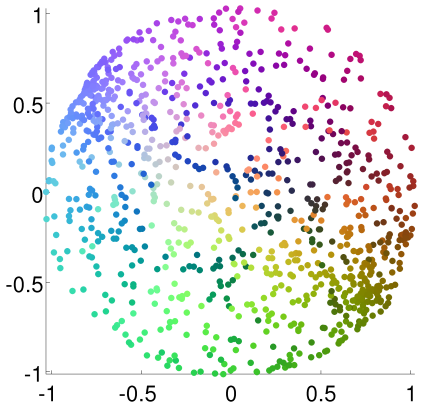
The idea presented by Venna and Kaski was to assign a specific cost to both types of errors and then minimise a combination of them. The new cost criteria inspired a new method called Neighbour Retrieval Visualiser (NeRV). Instead of minimising the combination of precision and recall directly, a more sophisticated method was developed to take into account grades of relevance as well as the rank order of the neighbours.

In the following paper [47], a probabilistic model was introduced. Developing the idea from Stochastic Neighbour Embedding (SNE) by Hinton and Roweis [16], Venna et al. defined a probability distribution over the neighbour points. The probabilistic model of retrieval defines a distribution $q_{j|i}$ that can be interpreted as the probability of a user labeling point j as neighbour of i in the output space. The distribution is defined to be

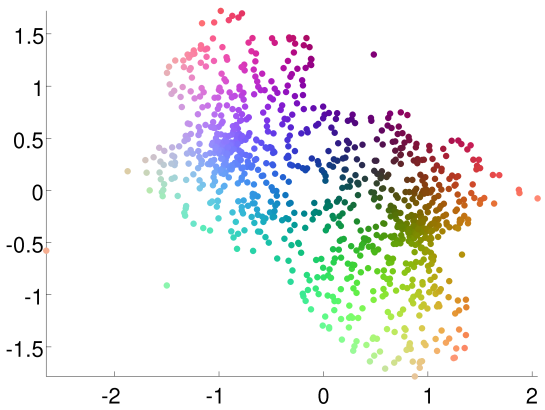
$$q_{j|i} = \frac{\exp(-\frac{||y_i - y_j||^2}{\sigma_i^2})}{\sum_{k \neq i} \exp(-\frac{||y_i - y_j||^2}{\sigma_i^2})}, \tag{3.8}$$



(a) Data on the surface of a three-dimensional sphere.



(b) The data projected in two dimensions by squashing the sphere flat.



(c) The data projected in two dimensions by cutting the sphere open.

Figure 3.3: A three dimensional sphere and two ways of projecting the surface in two dimensions.

where y_k is the location of sample k in the output space. $q_{j|i}$ is clearly Gaussian meaning that the probability of j being a neighbour of i decreases exponentially while their mutual distance increases.

The probabilistic model of relevance defines a probability distribution $p_{j|i}$ over the

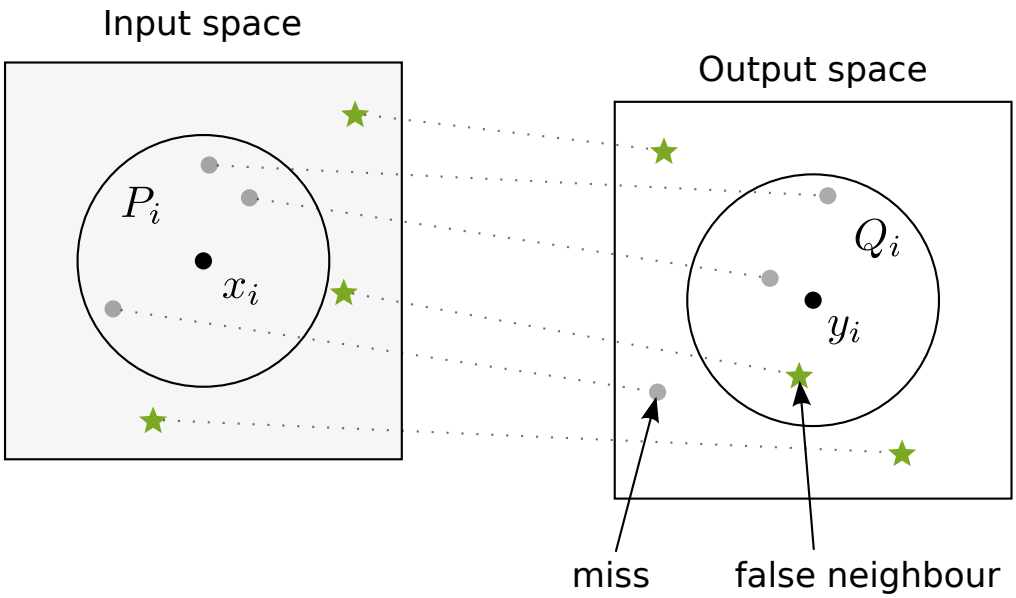


Figure 3.4: Visulising similarity relationships results usually in two kinds of errors: false neighbours and misses. Figure modified from [47].

neighbours in the input space. Analogously to $q_{j|i}$, the distribution becomes

$$p_{j|i} = \frac{\exp(-\frac{d(x_i,x_j)^2}{\sigma_i^2})}{\sum_{k\neq i}\exp(-\frac{d(x_i,x_j)^2}{\sigma_i^2})}, \tag{3.9}$$

where $d(\cdot,\cdot)$ represents the difference function used to calculate the dissimilarities in the original space and x_k is the location of sample k in the input space. The scaling parameter σ_i^2 present in both equations (3.8) and (3.9) controls how quickly the probabilities fall with increasing distance. This value was fixed such that the entropy of the distributions becomes $\log k$, where k is the number of relevant neighbours defined by the user.[47]

Now, the question is, how to compare the neighbours retrieved from the visualisation and the original neighbours in the input space. The neighbourhoods are described in terms of probability distributions, hence a natural way of comparing them is to use the Kullback-Leibler divergence:

$$D(p_i,q_i) = \sum_{j\neq i} p_{j|i}\log\frac{p_{j|i}}{q_{j|i}}. \tag{3.10}$$

If the distributions p_i and q_i are defined as in (3.8) and (3.9) it turns out that the Kullback-Leibler divergence defined here is a sort of generalisation of recall. Switching over the distributions in Equation (3.10) yields another measure, which can be considered a generalisation of precision. Venna et al. [47, Appendix A] show that for binary neighbourhoods the Kullback-Leibler divergences and the precision-recall measures become equivalent. Because of their similar nature, Venna et al. call $D(p_i,q_i)$ smoothed recall and $D(q_i,p_i)$ smoothed precision.

These loss functions can be used to measure the quality of a given visualisation. In addition, they are continuous and differentiable functions of the output visualisation coordinates, and can thus be used as optimisation criteria for an algorithm. However, many times visualisation is a compromise between precision and recall as was visualised in Figure 3.3. Thus, it's better to optimise a combination of the criteria and leave it to the user to decide which is more important: precision or recall. A relative cost of λ is assigned to misses and $(1-\lambda)$ to false positives yielding a total cost function:

$$\begin{aligned} E_{\text{NeRV}} &= \lambda \mathbb{E}_i[D(p_i, q_i)] + (1-\lambda)\mathbb{E}_i[D(q_i, p_i)] \\ &= \lambda \sum_i \sum_{j \neq i} p_{j|i} \log \frac{p_{j|i}}{q_{j|i}} + (1-\lambda) \sum_i \sum_{j \neq i} q_{j|i} \log \frac{q_{j|i}}{p_{j|i}}. \end{aligned} \tag{3.11}$$

Since the cost function above describes the visualisation task as a neighbour retrieval problem, the method that optimises (3.11) is called Neighbour Retrieval Visualiser.

By varying the parameter $\lambda \in [0, 1]$ the user can decide on which one of the two: precision or recall the method should focus. Setting $\lambda = 1$ means recall is maximised and the method behaves equivalently to the SNE of Hinton and Roweis [16]. Conversely, if $\lambda = 0$ the method tries to avoid false positives and thus maximises precision. Returning to the example in Figure 3.3, the flattened sphere in Figure 3.3b is produced when NeRV is applied to the original 3-D dimensional data of 3.3a and λ is set to 1. The sphere that has been cut open in Figure 3.3c is the result of applying NeRV with parameter $\lambda = 0$.

Given some value λ, the cost function may be used as optimisation criteria for the algorithm. Venna et al. [47] use conjugate gradient in order to find the optimal solution. The parameter σ_i is tuned during an initialisation phase to speed up convergence and avoid local minima. After the initialisation twenty conjugate gradient steps are performed to obtain the final solution. The computational complexity of the algorithm is $O(dn^2)$.

The cost function in (3.11) can be compared to the corresponding Stress function (3.4) of MDS. MDS attempts to map the original dissimilarities such that the resulting distances in the output space resemble them as closely as possible. Instead, NeRV models the original data as a set of probabilistic neighbourhoods and then tries to map the points so that the probability distribution is preserved as well as possible.

In NeRV, much of the attention is put to the form of the cost function while MDS emphasises more the relationship between the dissimilarities and distances. However, the question of how to obtain the dissimilarities remains the same. In NeRV, the choice of the original distance function $d(x_i, x_j)$ in (3.9) is left to the user. Again no additional transformations need to be done such as in metric and nonmetric MDS. Both methods can be used in the case when no dissimilarity information is available and only the rank order of the data points is known. In such a case, the NeRV equations (3.8) and (3.9) are modified such that $p_{j|i}$ and $q_{j|i}$ are replaced with ranks.

NeRV is a nonlinear dimensionality reduction method especially designed for visualisation purposes. It aims to minimise a special cost function that is a combination of generalised versions of precision and recall. NeRV is based on the idea that visualisation is inherently an information retrieval problem. It answers the question: how can the neighbourhood of a given point be retrieved as well as possible based on the visualisation? Design choices can be made by varying the value of the parameter λ. Setting $\lambda = 0$ makes NeRV avoid only false neighbours and not care about misses. Whereas setting $\lambda = 1$

causes the algorithm to minimise the number of misses and ignore false neighbours. A value between $[0, 1]$ will result in a compromise between these two extremes.

3.2.3 Comparison

Why would one choose some visualisation method over another one? The methods presented here are defined in different ways and thus take a slightly different perspective to visualisation and dimensionality reduction. The answer depends on what the overall purpose of the visualisation is. In this case, the resulting visualisation is supposedly used in a peer support service, hence it is interesting to measure the ability to retrieve a neighbourhood structure based on a visualisation. Aspects like computational complexity, availability and ease of implementation should be considered as well since the method is supposed to be used in a real-life application.

Three real life data sets were chosen for comparing the visualisation performance of NeRV and MDS. Experiments were conducted on two wellbeing data sets and one common benchmark data set.

The *Wine Data Set* is a widely used benchmark from the UCI Machine Learning Repository by Frank and Asuncion [11]. The data consists of 178 wines on which 13 different chemical measurements have been performed. The wines are grown in the same region in Italy, but they're derived from three different cultivars, which can be interpreted as classes in the machine learning perspective. The 13 attributes are: alcohol, malic acid, ash, alcalinity of ash, magnesium, total phenols, flavanoids, nonflavanoid phenols, proanthocyanins, color intensity, hue, OD280/OD315 of diluted wines and proline. All the attributes are continuous integer or real valued. For the visualisation, the data was normalised such that each variable had zero mean and unit variance.

The *Stress Data Set* is a result of a stress survey conducted at the WIC-2011 Wellbeing Innovation Camp [4] in 2011 which was arranged as a part of the VirtualCoach Project. The 43 respondents were students who participated in the WIC-2011. The survey consisted of 33 multiple-choice questions from which 17 are about experiencing stress and the remaining 25 consider coping and stress relieving methods. Answers to the questions were given on a Likert-scale from 1 to 5. The survey included also open text questions, which were ignored for now.

There are 43 answers in the data set and a total of 33 variables. The first 17 variables were chosen based on some earlier work on the data and also because the remaining 25 variables contain missing values. The values of the attributes are discrete ranging from 1 to 5. The full list of questions is found in the Appendix A.

The third data set is an extensive collection of health related measurements from the Sport Institute of Finland at Vierumäki. The complete data set consists of more than 100000 measurements from both males and females gathered during the years 1998–2009. A subset of the data set beginning from the year 2006 was used because during that time the measurement protocol remained the same. Each sample consists of 10 measurements: age, level of fitness, body mass index (BMI), percentage of fat and six mobility measurements including legs, arms, abdomen, shoulders, sides and hips. The attributes are continuous integer and real valued. The subset still contained over 30000 samples, so to reduce the computational load a random sample of 2000 data points from each gender was chosen for analysis.

Three pairs of goodness measures were used to compare the visualisation performance of the methods. The first pair is *mean smoothed precision* and *mean smoothed recall* as defined in Sub-section 3.2.2. The scale of the neighbourhood was set to 20 relevant neighbours. Standard *mean precision* and *mean recall* were measured so that the 20 nearest neighbours of a point in the original data were chosen as the set of relevant neighbours. Then the number of true neighbours, false neighbours and misses were calculated varying the number of neighbours from 1 to 100 in the visualisation space. This resulted in curves of mean precision and mean recall for each method. The third pair of measures was a variant of the mean smoothed precision and mean smoothed recall, in which the distances in (3.9) and (3.8) are replaced with ranks so that the nearest neighbour has the distance 1, the second nearest 2 and so on. These *mean rank-based smoothed precision* and *mean rank-based smoothed* recall provide an alternative way of comparing the methods. Again the number of relevant neighbours was set to 20. All the above choices of parameters were as suggested in [47].

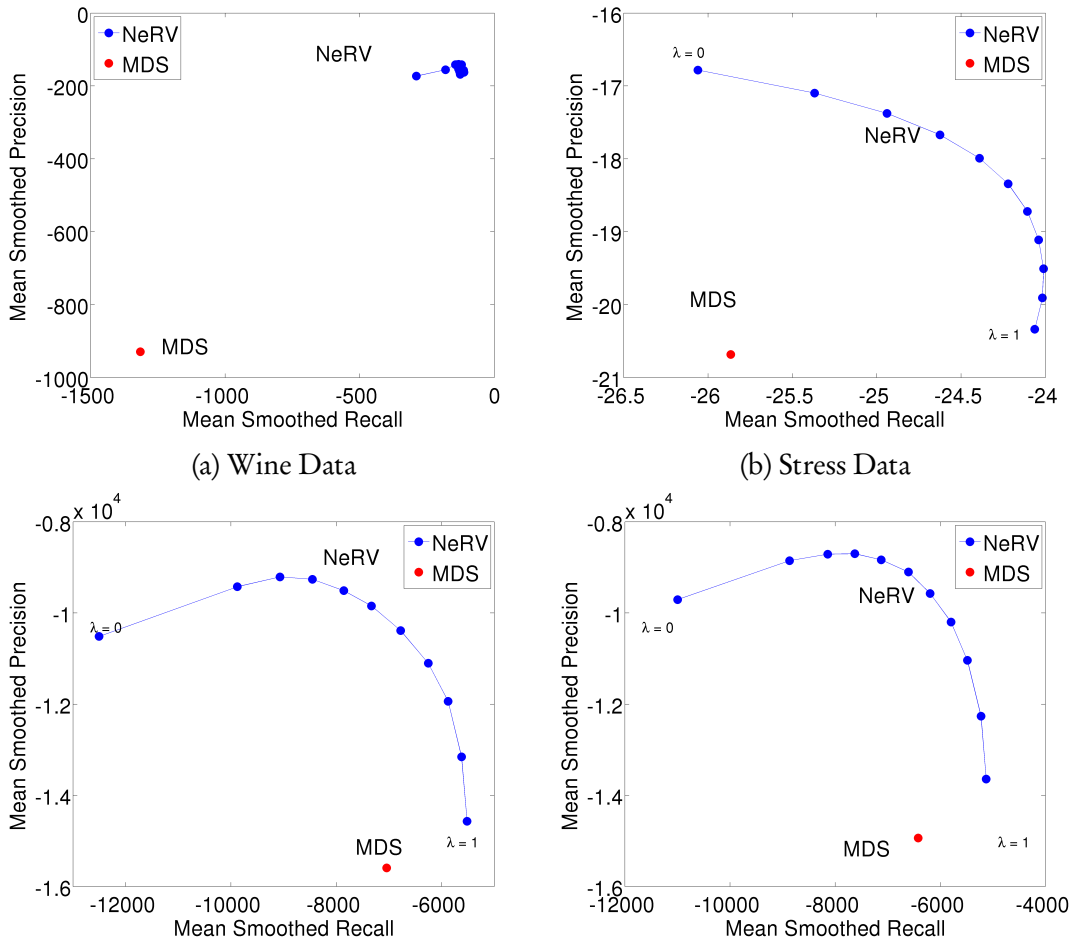
Figure 3.5 shows the curves for the mean smoothed precision and mean smoothed recall. The performance of NeRV has been plotted for several values of λ producing a curve. The results for each of the plots are scaled to maintain visual consistency showing the best performing methods in the top right corner. NeRV performs better on all three datasets, which is not surprising as NeRV is specifically designed to optimise these criteria. It is also clearly visible from Figures 3.5b, 3.5c and 3.5d how the parameter λ affects the result. When $\lambda = 0$ NeRV maximises only smoothed precision, which results in a lower value of smoothed recall on the upper left corner of the figure. Setting $\lambda = 1$ results in an increased smoothed recall, but at the same time smoothed precision decreases in the lower right corner.

Standard mean precision and mean recall was also calculated for the three data sets. The results are in Figure 3.6. The curve is shown for a single value of λ chosen so that it maximises the F-measure: $2(P \cdot R)/(P + R)$, where P and R are the mean rank-based smoothed precision and mean rank-based smoothed recall as suggested in [47]. MDS performs as well or slightly better than NeRV in terms of mean precision and mean recall. NeRV yields higher precision when considering smaller neighbourhoods especially in 3.6c and 3.6d. In Figure 3.6b, MDS performs clearly better, although the difference is small. Both perform well on the stress data.

Finally, the results of measuring mean rank-based smoothed precision and mean rank-based smoothed recall are shown in Figure 3.7. The NeRV curve is again parametrised by λ. The curves resemble the ones in Figure 3.5 with the exception of the stress data in Figure 3.7b. Although NeRV achieves clearly higher values on on the Stress data in terms of smoothed precision and smoothed recall, MDS is better when comparing the rank-based measures.

It fair to say that with suitable values of λ NeRV achieves higher values of precision and recall. Values of lambda should be chosen between 0.2 – 0.8 to achieve the best performance. A comprehensive treatment and comparison of the methods would be beyond the scope of this thesis. The experiments presented here demonstrate that both MDS and NeRV can be used to visualise complex multidimensional data. More extensive experiments documenting the visualisation performance of NeRV, MDS and several other visualisation methods can be found in [47].

In terms of computational complexity, both methods are of the order $O(n^2)$. Venna



(a) Wine Data

(b) Stress Data

(c) Fitness Data Men

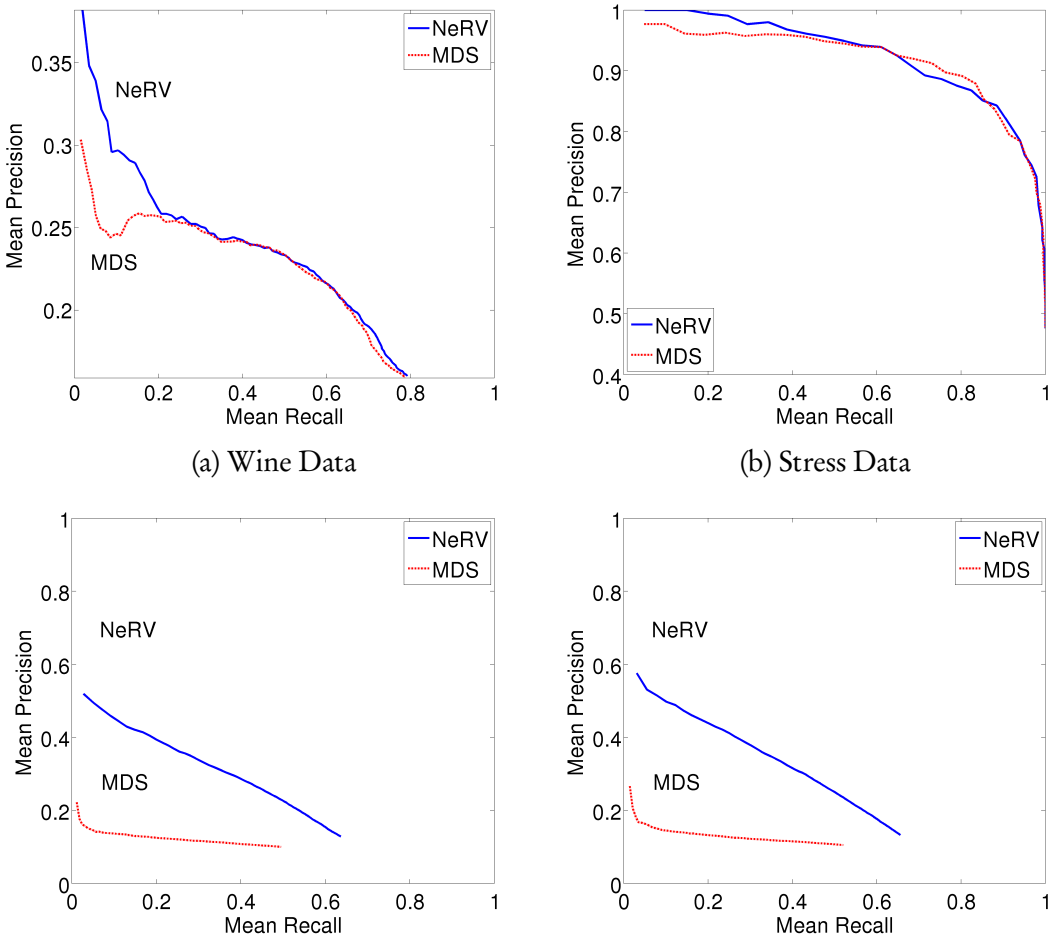
(d) Fitness Data Women

Figure 3.5: Mean smoothed precision and mean smoothed recall plotted for the three data sets. Actually -1·(mean smoothed precision) and -1·(mean smoothed precision) were plotted to maintain consistency with the other plotted measures. The best performing method appears in the top right corner.

et al. [47] report that the computational complexity of optimising the NeRV cost function (3.11) is $O(dn^2)$, where n is the number of data points and d is the dimension of the projection. Similarly, in MDS, pair-wise comparisons between all the samples need to be made during each iteration resulting in a computational complexity of the order of $O(n^2)$.

Several implementations of MDS exist, and variants of it are included in many statistical software systems. NeRV is a recent algorithm and there exists only one C++ implementation found in [1]. Although having roughly the same order of computational complexity, the current implementation of NeRV runs notably slower than the MATLAB implementation of MDS. Implementations of both algorithms are freely available and relatively straightforward to use.

Nonlinear projection methods such as MDS and NeRV do not provide an explicit mapping function, so that the coordinates need to be recalculated every time a new data



(a) Wine Data

(b) Stress Data

(c) Fitness Data Men

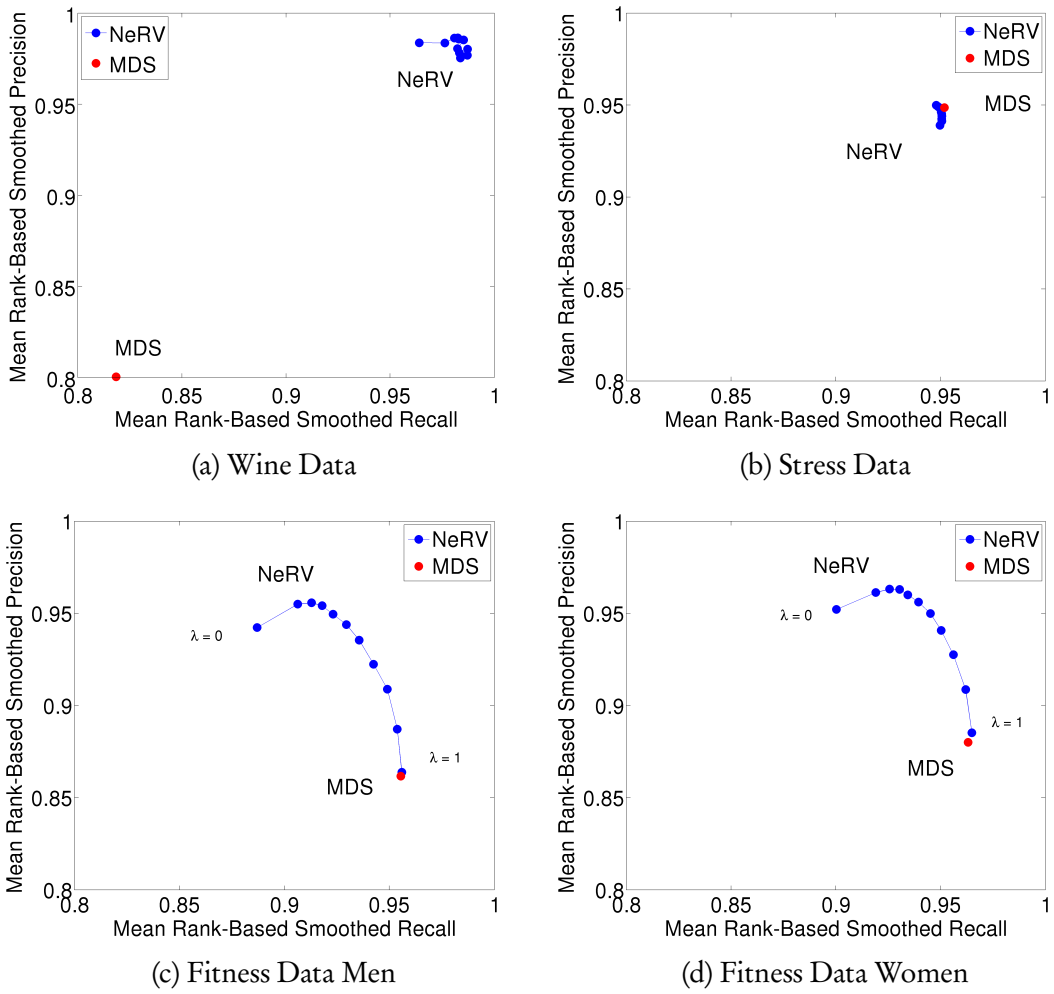
(d) Fitness Data Women

Figure 3.6: Mean precision and mean recall plotted for the three data sets. The best performing method appears in the top right corner.

point is introduced. This could be a major disadvantage when considering a real time online application, which might have several concurrent users and new data coming in all the time. This is not the case with the prototype vector based methods that are presented in the following section.

3.3 Topographic Map Methods

In this section, two more methods for visualisation are presented. The Self-Organising Map (SOM) is a well established methodology originally conceived by Kohonen [25]. A probabilistic alternative to SOM called Generative Topographic Mapping (GTM) was developed in [7]. Both methods are characterised by the existence of a regular array of nodes which the original data points are mapped into.



(a) Wine Data

(b) Stress Data

(c) Fitness Data Men

(d) Fitness Data Women

Figure 3.7: Mean rank-based smoothed precision and mean rank-based smoothed recall plotted for the three data sets. Actually 1-(mean rank based smoothed precision) and 1-(mean rank based smoothed recall) were plotted to maintain visual consistency with the other measures. The best performing method appears in the top right corner.

3.3.1 Self-Organising Maps

Self-Organising Maps are widely used in nonlinear dimensionality reduction, vector quantisation and visualisation in many fields of research. A typical SOM consists of a two-dimensional regular array of nodes, often called map units or prototype vectors. An explicit topological neighbourhood is defined around the nodes like the rectangular neighbourhoods in Figure 3.8, where $N_i(t)$ is the neighbourhood set of the map unit i at time t.

Each map unit is associated with a reference vector m_i in the high-dimensional input space. Each data point is mapped into the map unit whose reference vector it most closely resembles. Thus, a map unit represents similar data points and serves as a prototype for them in the output space. An example of fitting a SOM to artificial data is presented in Figure 3.9.

Introducing some notation, let $\mathbf{X}$ be the m-dimensional data matrix with n samples

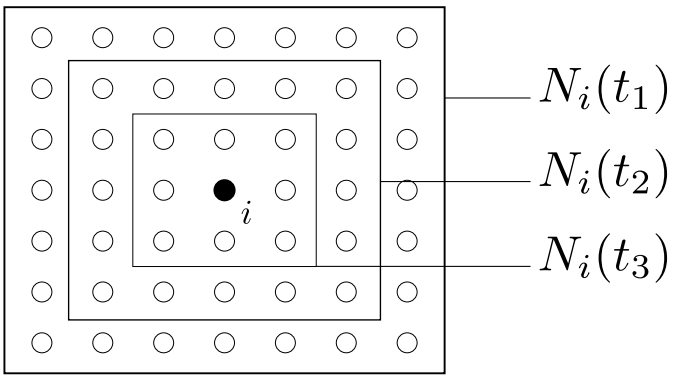


Figure 3.8: Examples of topological neighbourhoods at times ($t_1 < t_2 < t_3$).

and $\mathbf{m}_i$ be the m-dimensional reference vector associated with map unit i. The reference vector $\mathbf{m}_c$ that most closely matches a given data vector $\mathbf{x}_t$ is found by computing the distance:

$$d(\mathbf{x}_t, \mathbf{m}_c) = \min_i \{d(\mathbf{x}_t, \mathbf{m}_i)\}. \tag{3.12}$$

Usually the Euclidean distance is used, but some other metric could be used as well, as demonstrated in [26].

After finding out the best matching unit (BMU) the reference vectors are updated so that $\mathbf{m}_i$ is shifted towards the data sample $\mathbf{x}_t$. In addition, the neighbours of the BMU are shifted towards the data sample. This peculiarity makes all the difference and is responsible for the self-organising nature of SOM. Because whole neighbourhoods are learning together it imposes a natural ordering in the resulting visualisation. The update can be put more formally:

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + h_{ci}(t)[\mathbf{x}(t) - \mathbf{m}_i(t)], \tag{3.13}$$

where h_{ci} refers to the neighbourhood function defined over the neighbours of c. The neighbourhood function can be discrete so that $h_{ci} = \alpha(t)$ when $i \in N_c$ and $h_{ci} = 0$ when $i \notin N_c$, where the sets N_c can be time dependent as in Figure 3.8. However, a smooth kernel is often preferred. Denoting the coordinates of units c and i with vectors $\mathbf{r}_c$ and $\mathbf{r}_i$ the neighbourhood function becomes:

$$h_{ci}(t) = \alpha(t) \cdot \exp\left(-\frac{||\mathbf{r}_i - \mathbf{r}_c||^2}{\sigma^2(t)}\right). \tag{3.14}$$

$\alpha(t)$ is a scalar learning-rate, and σ the width factor of the kernel. Both are decreased over time to ensure convergence. Initialisation of the reference vectors $\mathbf{m}_i(0)$ needs to be done carefully as well.

The size of the SOM or the number of map units needs to be chosen by the user. The choice depends heavily on the purpose of the visualisation, but some validation of map sizes is recommended. The mathematical theory behind SOM is somewhat complicated and there is no objective function that measures whether a particular map is optimal for a given task. For validation there are two commonly used error measures: the quantisation error and the topological error. The former measures how much the original data vector

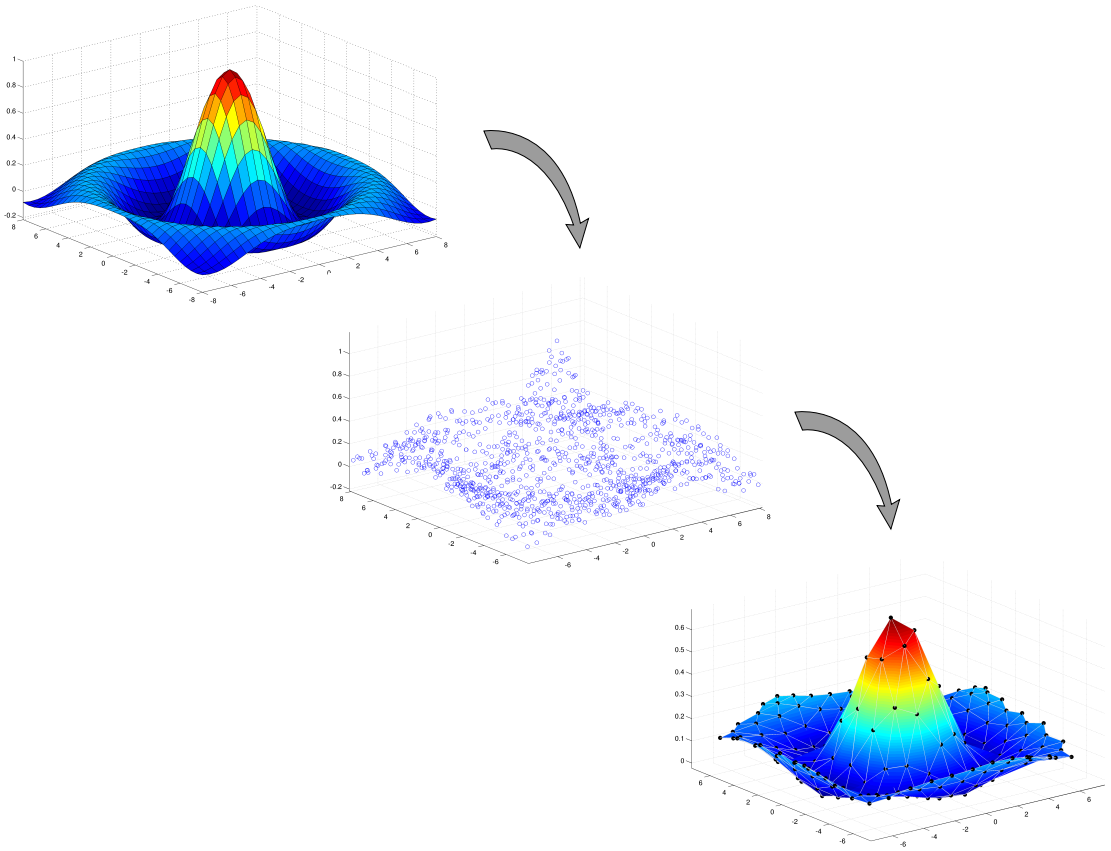


Figure 3.9: Fitting a SOM for artificial data. The original function is on the left. A random sample from the function with added noise in the middle. The resulting SOM on the right. The map units have been displayed with black dots. The locations of the map units on the original 3-d space are determined with their corresponding reference vectors.

$\mathbf{x}$ differs from the best matching unit and the latter is the fraction of data points whose two closest map units are not neighbours in the topology. These measures are inversely related so that when the map size increases usually the quantisation error goes down and the topological error increases. Kaski and Lagus [21] have developed an additional measure called combined error, which is the quantisation error together with the distance from the BMU to the second-best-matching unit along the path on the topology.

Choosing the correct number of map units is not always straightforward similarly as choosing the correct number of clusters in unsupervised clustering problems. The quality of the map can also be inspected visually by looking at the distribution of the values of the variables. These distributions are called the component planes and they are used when the results from the SOM are interpreted.

To conclude, the SOM is an unsupervised dimensionality reduction method, useful in making two-dimensional maps of high-dimensional data. By compressing the data into prototypes SOM also performs a sort of abstraction and can be thought of finding latent features of the data.

3.3.2 Generative Topographic Mapping

The problem of visualising high-dimensional data can be examined in another way with the help of latent variable models. The idea is to represent the probability density of the data in terms of a smaller number of latent variables. Generative Topographic Mapping is a latent variable model that provides a nonlinear mapping from data space to a lower dimensional latent or output space. It was designed by Bishop et al. [7] to overcome some of the limitations present in the SOM. Unlike the SOM the GTM algorithm defines an explicit probability distribution over the data and thus can use its log likelihood function as an objective cost criterion for evaluation purposes.

GTM serves as a visualisation method mapping data from the m-dimensional space to a lower dimensional visualisation space. However, the model is defined backwards so that the mapping is actually from the latent space into the data space. Visualisation is then performed by computing the posterior distribution of the data in latent space by using Bayes' theorem. [7]

Let $p(\mathbf{x})$ be the distribution of the data in a m-dimensional space $\mathbf{x} = (x_1, \ldots, x_m)$. The GTM tries to find a representation for this distribution in terms of d latent variables $\mathbf{u} = (u_1, \ldots, u_d)$. For visualisation, d is usually 2 or 3. The mapping is achieved by considering a function $\mathbf{y}(\mathbf{u}; \mathbf{W})$, where $\mathbf{W}$ is a matrix containing the parameters of the model. The transformation maps points u in latent space to points in the data space. This is illustrated schematically in Figure 3.10.

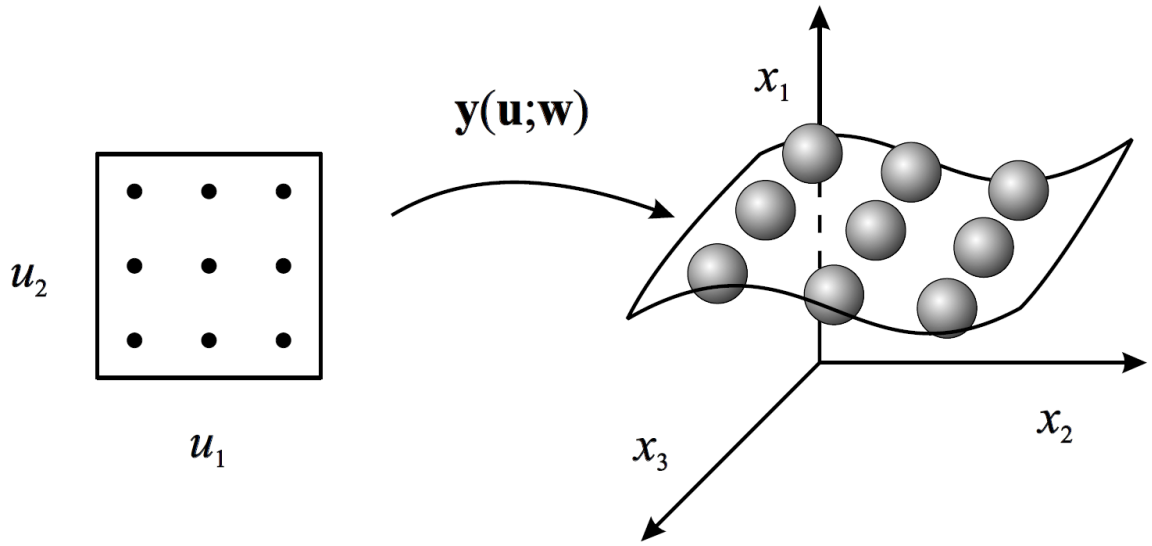


Figure 3.10: The latent variable model of the GTM illustrated schematically as in [7]. A regular grid of nodes in the latent space is visualised on the left-hand side. Each node $\mathbf{u}_i$ is mapped to the corresponding unit $\mathbf{m}_i = \mathbf{y}(\mathbf{u}_i; \mathbf{W})$, and forms the centre of an isotropic Gaussian distribution in data space.

Similarly to the SOM a regular array of nodes $\mathbf{u}_i$, $i = 1, \ldots K$, is defined in the latent space. In addition, a set of M nonlinear radial basis functions $\phi(\mathbf{u}) = \{\phi_j(\mathbf{u})\}$, where $j = 1, \ldots, M$, is introduced. Using these the mapping from latent space to data space is given by:

$$\mathbf{m}_i = \mathbf{W}\phi(\mathbf{u}_i), \tag{3.15}$$

where $\mathbf{m}_i$ are the corresponding reference vectors in data space and $\mathbf{W}$ is a $m \times M$ matrix of weight parameters. The model is formed around these reference vectors so that each of them serves as the centre of an isotropic Gaussian distribution in data space:

$$p(\mathbf{x}|i) = \left(\frac{\beta}{2\pi}\right)^{D/2} \exp\left\{-\frac{\beta}{2}||\mathbf{m}_i - \mathbf{x}||^2\right\}, \tag{3.16}$$

where β is the inverse variance of the Gaussian. The above Gaussian distribution also accounts for a noise model considering the usual occasion that the data does not lie exactly on the lower-dimensional manifold.

The probability density of the GTM model is obtained by summing over all of the K Gaussian components yielding:

$$p(\mathbf{x}|\mathbf{W},\beta) = \sum_{i=1}^{K} P(i)p(\mathbf{x}|i) = \sum_{i=1}^{K} \frac{1}{K}\left(\frac{\beta}{2\pi}\right)^{D/2} \exp\left\{-\frac{\beta}{2}||\mathbf{m}_i - \mathbf{x}||^2\right\}. \tag{3.17}$$

The prior probabilities $P(i)$ are constant and set to $1/K$. The resulting model can be regarded as a constrained mixture of Gaussians. The model parameters $\mathbf{W}$ and β are adapted through learning, but the number of basis functions M and the number of nodes K in the latent space have to be determined by the user.

The parameters $\mathbf{W}$ and β are fitted to a data set $\{\mathbf{x}_t\}$, where $t = 1,\ldots,n$ by maximum likelihood. The log likelihood function is given by:

$$\mathcal{L}(\mathbf{W},\beta) = \sum_{t=1}^{n} \ln p(\mathbf{x}_t|\mathbf{W},\beta). \tag{3.18}$$

The log likelihood can be maximised using the EM-algorithm and the function can be evaluated during the algorithm to monitor the convergence.

When the GTM is used in visualisation, the mapping from latent space to data space needs to be inverted using the Bayes' theorem. The posterior of a node or the responsibility of a node for explaining $\mathbf{x}_t$ is given by:

$$\gamma(t)_i = p(\mathbf{u}_i|\mathbf{x}_t,\mathbf{W},\beta) = \frac{p(\mathbf{x}_t|\mathbf{u}_i,\mathbf{W},\beta)}{\sum_{i=1}^{K} p(\mathbf{x}_t|\mathbf{u}_i,\mathbf{W},\beta)}. \tag{3.19}$$

However, it is difficult to visualise the full posteriors of all the data points, so the posterior must be summarised for example by its mean:

$$\langle \mathbf{x}_t|\mathbf{u}_i,\mathbf{W},\beta\rangle = \sum_{i=1}^{K} \mathbf{x}_t \; p(\mathbf{u}_i|\mathbf{x}_t,\mathbf{W},\beta) \tag{3.20}$$

or by its mode:

$$i^{\max} = \arg\max_i \; p(\mathbf{u}_i|\mathbf{x}_t,\mathbf{W},\beta). \tag{3.21}$$

The posterior distribution can be multi-modal so that the locations of the mean and the mode might differ significantly. Thus, careful attention should be put to the shapes of the posteriors when visualising complete data sets.

3.3.3 Comparison

Despite the rivalry between the SOM and the GTM (see, e.g., [20, p. 28–29] and [7, p. 12]) the methods are compared from the point of view of visualising neighbourhoods. The question is: which method is more suitable for visualising proximities? The topographic mapping methods were compared using the same framework introduced in the previous section.

The same three data sets: the wine data, the stress data and the two fitness data sets were used in comparison. The same measures of goodness were also applied to each of the methods in each of the data sets.

The calculations were performed so that first the number of latent map units was determined by training the SOM with different map sizes ranging from 10 to 100 for the stress data, from 10 to 600 for the wine and from 400 to 4000 for the fitness data. A 10-fold cross-validation was performed and the mean combined error was plotted for all the sizes. The final map size was chosen from the error plot subjectively. The same number of map units was then used in training the GTM. In addition, the number of RBF centres for the GTM was chosen to be 9 for the stress data and 16 for the two other data sets without optimising it further.

The methods were trained using the chosen number of map units and then each data point was mapped into the best matching unit. In the GTM, the mode of the posterior distribution was used as the location in the visualisation space. Then distances between the points were calculated and neighbours were determined based on those distances. Some points were mapped into the same map units yielding 0 distances between them.

Mean smoothed precision and mean smoothed recall were calculated for only one map size resulting in a single measurement for both methods. Standard mean precision and mean recall were calculated for the three data sets in the same way as before. The 20 nearest neighbours of a point in the original data were chosen as the set of relevant neighbours. Then the number of true neighbours, false neighbours and misses were calculated varying the number of neighbours from 1 to 100 in the visualisation space. This resulted in curves of mean precision and mean recall for each method. Mean rank-based smoothed recall and smoothed precision were also compared and calculated for both methods.

Table 3.2: Wine Data

	Mean smoothed precision	Mean smoothed recall
SOM	-227.2	-235.1
GTM	-227.1	-380.9
	Mean rank-based smoothed precision	Mean rank-based smoothed recall
SOM	0.97	0.97
GTM	0.95	0.90

Assessing the results in tables 3.2, 3.3, 3.4 and 3.5 it can be stated that the order of magnitude of the values is similar to those achieved by NeRV and MDS. The measures are consistently worse than those achieved by NeRV and MDS but this is natural as these measures of goodness are not designed for evaluating these kind of topographic mapping methods. The same framework for evaluating the visualisation performance was used for consistency.

Table 3.3: Stress Data

	Mean smoothed precision	Mean smoothed recall
SOM	-29.1	-50.4
GTM	-25.7	-53.4
	Mean rank-based smoothed precision	Mean rank-based smoothed recall
SOM	0.89	0.89
GTM	0.89	0.90

Table 3.4: Fitness Data Women

	Mean smoothed precision	Mean smoothed recall
SOM	-10327	-10535
GTM	-11873	-12432
	Mean rank-based smoothed precision	Mean rank-based smoothed recall
SOM	0.90	0.96
GTM	0.88	0.94

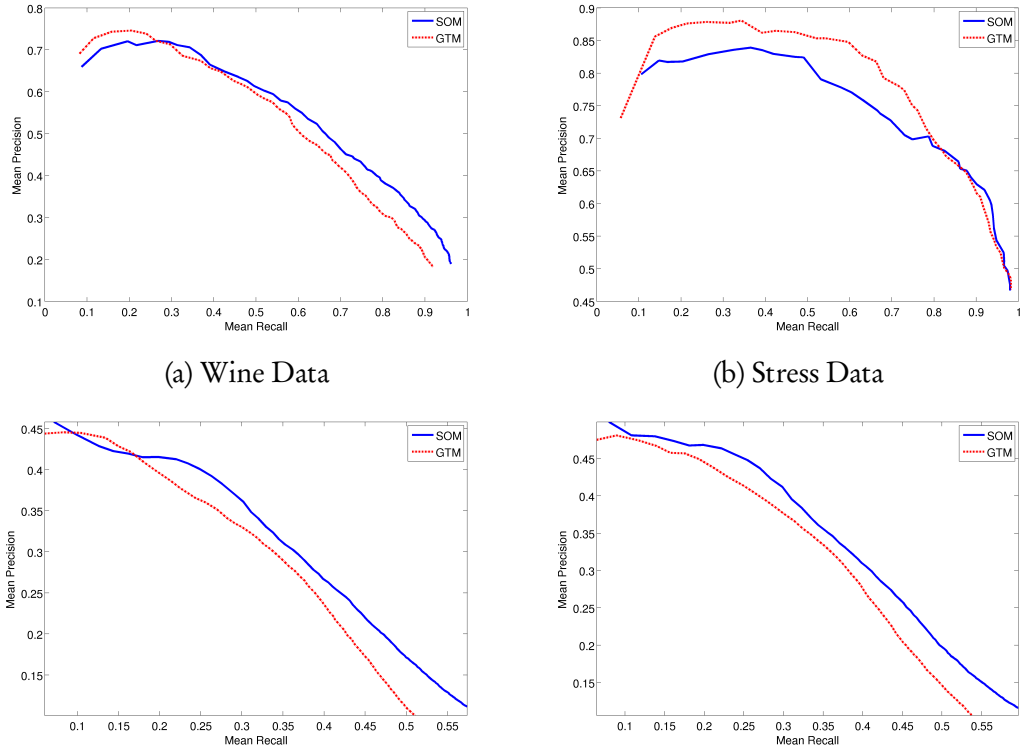
Table 3.5: Fitness Data Men

	Mean smoothed precision	Mean smoothed recall
SOM	-10894	-12101
GTM	-12585	-13366
	Mean rank-based smoothed precision	Mean rank-based smoothed recall
SOM	0.88	0.95
GTM	0.87	0.93

The resulting curves for calculating mean precision and mean recall can be found in Figure 3.11. Surprisingly theses curves do share some similarity to the ones in Figure 3.6c. It also seems that the SOM and the GTM perform significantly better than the NeRV and the MDS. On the stress data, and both the fitness data sets, the SOM and the GTM perform quite poorly.

The most computationally loaded task in both the SOM and the GTM is the evaluation of the Euclidean distance between the data points and the reference vectors. Since both methods perform this same task at each iteration Bishop et al. [7] argue that the relative computational cost of both methods is approximately the same. Training the SOM on the women's fitness data with 2000 data points took 4.4 seconds while the GTM took 10.8 sec. Using 20000 samples the SOM took 16.2 seconds to converge while the GTM took 266.8 seconds.

The SOM is available for use as an extensive MATLAB Toolbox (http://www.cis.hut.fi/projects/somtoolbox/ [48]) or as a C implementation called som_pak [24]. The som_pak is especially useful for training large maps as it runs significantly faster than the MATLAB version. The GTM is included in the Netlab package [2] for use in MATLAB provided by the nonlinearity and Complexity Research Group from the Aston University. The toolbox offers several useful functions for training and using the GTM algorithm. However, a C implementation would be necessary for practical purposes as the MATLAB Toolbox can result in slow performance.



(a) Wine Data

(b) Stress Data

(c) Fitness Data Men

(d) Fitness Data Women

Figure 3.11: Mean precision and mean recall plotted for the three data sets.

Unlike the MDS and the NeRV the topographic mapping methods rely on a set of map units or nodes which can be used to map new incoming data points to the map without recalculating the whole model. In addition, the methods can be trained iteratively so that each new data point can be used to train the map little by little. This only changes the best matching unit and its neighbours. The downside is that the number of map units is a parameter that needs to be chosen by the user. Model selection is an art in itself and choosing the right parameters is rarely straightforward.

Returning to the question of on what grounds should one make a decision when choosing a method for visualisation. The four methods have now been compared using the framework designed by Venna and Kaski [46] to measure the performance of retrieving neighbours based on the visualisation. Since this framework was mainly designed to demonstrate the performance of NeRV, the measures are not as well suited for evaluating the topographic map methods. On most performance measures and most data sets the NeRV performs better or at least as well as the MDS. The results from the comparison between the SOM and the GTM are not as decisive probably because of the aforementioned reasons.

To test the implementation of each method in an actual visualisation task, another type of comparison was conducted. The visualisations were compared in a user study presented in chapter 5. Before going to that it is necessary to give an idea of how these methods could be implemented in a real world application. An example case of using visualisation to find peers is explained in the next chapter.

Chapter 4

Visualising Peers

The underlying fundamental assumptions of this thesis are:

- data analysis methods can help to find peers in the web,
- visualising the peer relationships is useful.

The fundamental questions following these assumptions are: what should this kind of a peer support application look like and how should the visualisations be done? This Chapter attempts to answer these questions. The outline and purpose of an example peer support application is presented in Section 4.1. The visualisation methods presented in the previous Chapter are implemented in the application and example visualisations are provided.

4.1 A Peer Support Application

The Internet allows for like minded and similar people (see Section 2.2 for a detailed discussion on similarity) to be in contact no matter how far away they are located from each other. The ability to communicate with people far away opens up vast possibilities for receiving support as well as for offering help to others.

In this section, an example of a peer support service is sketched. There are several existing web-based peer support services and the purpose of this thesis is not to suggest another such service. The aim is to analyse the design of such a service through the use of an example and possibly provide recommendations to consider when visualising peers. The idea of designing a user interface based on a visualisation is presented, e.g., in [17]. Another project incorporating the SOM with a visual user interface is the WEBSOM (see [22],[32]).

The hypothetical peer support service is called the Stress Map and it is based on the Stress Questionnaire introduced in Chapter 3. The Stress Map is intended to be an example of a web-based peer support service, where people with different backgrounds gather to share their experiences on stress and coping and to help each other overcome the problems they are dealing with. The idea is that the service helps the users to find their peers by visualising them on a 2-dimensional map where people "who are similar to them" are mapped close by and people less similar further away.

As the users enter the service, they are provided with a questionnaire that has several multiple-choice questions as well as open text questions. The purpose of the questionnaire is to ask the users about their ways of experiencing stress and how they cope with it. This is the initial input the users need to give before they can get any feedback from the service. A specific profile is formed from the answers to this questionnaire and this profile is used when mapping the users on the Stress Map. The questions on the questionnaire have to be well formed as the profile of each user is determined solely based on them. If the questions successfully capture the essential aspects of the experience of stress, then determining peers based on those questions is more credible. If, however, the questions seem irrelevant or misplaced, then the answers do not serve well in profiling users, and the similarity between people will seem vague at best. This is exactly the question discussed in Section 2.2. Choosing the questions is a case of choosing the correct set of features that represent the users in the best possible way. This is perhaps the most demanding part of designing such a service, and it is also very hard to evaluate the quality of any given questionnaire. In the end, it is the users who decide if a system seems plausible or not.

The questionnaire used in gathering the Stress data set was not especially designed for this kind of profiling, nor to test any hypothesis, but to survey the experience of stress of a group of students. The purpose of the Stress Map is to provide an example of a peer support service, so the suitability of this particular questionnaire is not important. The questions are found in the Appendix A. The 43 respondents of the Stress data set form the hypothetical first users of the Stress Map service. Their profiles consist of their answers to the Stress Questionnaire. After entering their answers to the service, they get a profile page showing their basic facts and their answers to the questions. Users can of course choose which of their personal information is shown to other users. If they wish, they can remain completely anonymous. Contacting and receiving help from other users requires some form of authentication so that mutual trust can be established.

As the name Stress Map suggests, the prominent component of the service is the map of peers. It serves as the principal search tool that helps to browse through the users of the service. The map is intended to give the users the power to find the kind of peers that they deem necessary. The map positions the users of the service on a plane according to their similarity. The users are presented with their own location on the map after which they can browse the peers located close or far or however they wish. When the users find a peer they find interesting, they can view the profile of the peer and as well as their answers to the questions. The peers may have written stories of their own experiences, or they might even have written hints or suggestions of what others could do to help themselves in the same situation. The users may also decide to contact a peer through the service and perhaps ask more or offer their own experiences. Preferably these conversations would be added to the service for others to read and learn, but it is entirely up to the peers to decide whether they will be made public or not.

To conclude, this rough description of the Stress Map service is supposed to give an idea of what the service is intended to do and how it is used. However, it makes little sense to try to imagine this kind of visualisation based service without any images. That is why the second contribution of this thesis is to produce examples of how to use the visualisation methods, presented in the previous Chapter, in actual visualisations. In the next section some example visualisations are provided to show how the service might look like for the end user.

4.2 Visualising Peers

Comparing visualisation methods with scientific measures of goodness is important, but looking at the resulting visualisations is more fun. In this section the methods presented in Chapter 3 are used in practice. The Stress data set was projected with the different methods and the resulting visualisations were used to design views of the Stress Map service.

The views were also designed so that a user study could be conducted later, to assess the advantages and disadvantages of the visualisations. For this reason both the MDS and NeRV were not used, because from the point of view of visualisations they produce identical results. Both produce a cloud of points in a 2-dimensional continuous space, the only difference being that the points are distributed a little differently. When the images were assessed for their quality of visualisation there was no reason to study MDS and NeRV separately, as the same remarks apply for both of them. NeRV was chosen for the visualisation because it has constantly better precision and recall and its performance can be altered with the use of the parameter λ.

For each of the methods NeRV, the SOM and the GTM two views were drawn. The first one shows the basic view of the service and the second represents the action of browsing or clicking a peer with the mouse. The colours of the dots represent a clustering of the data that was performed beforehand and separately from the visualisations. It was done as a preprocessing step to study if potential clusters could be found in the data. However, the data consists of only 43 data points so the clustering cannot be considered very definitive. The colours are drawn in the views to help to interpret and compare the methods.

The views produced with NeRV are found in Figures 4.1 and 4.2. The SOM views are in Figures 4.3 and 4.4. Finally the GTM-based views are found in Figures 4.5 and 4.6. An example of a profile view is in Figure 4.7.

The visualisations serve as a tool for better grasping the concepts discussed so far. To discuss and assess something, it is first necessary to see and get a feeling of the subject. The visualisations were made for this exact reason, to convey ideas and literally to visualise a peer support service. However they were designed based on an initial idea and some preliminary vision of the service and they serve only as an introduction. When developing a real world service with potential users, it is almost axiomatic to ask the users for feedback at nearly every stage of development. For this reason, a small user study was carried out to gather insight from people. The study and the results are presented in the next Chapter.

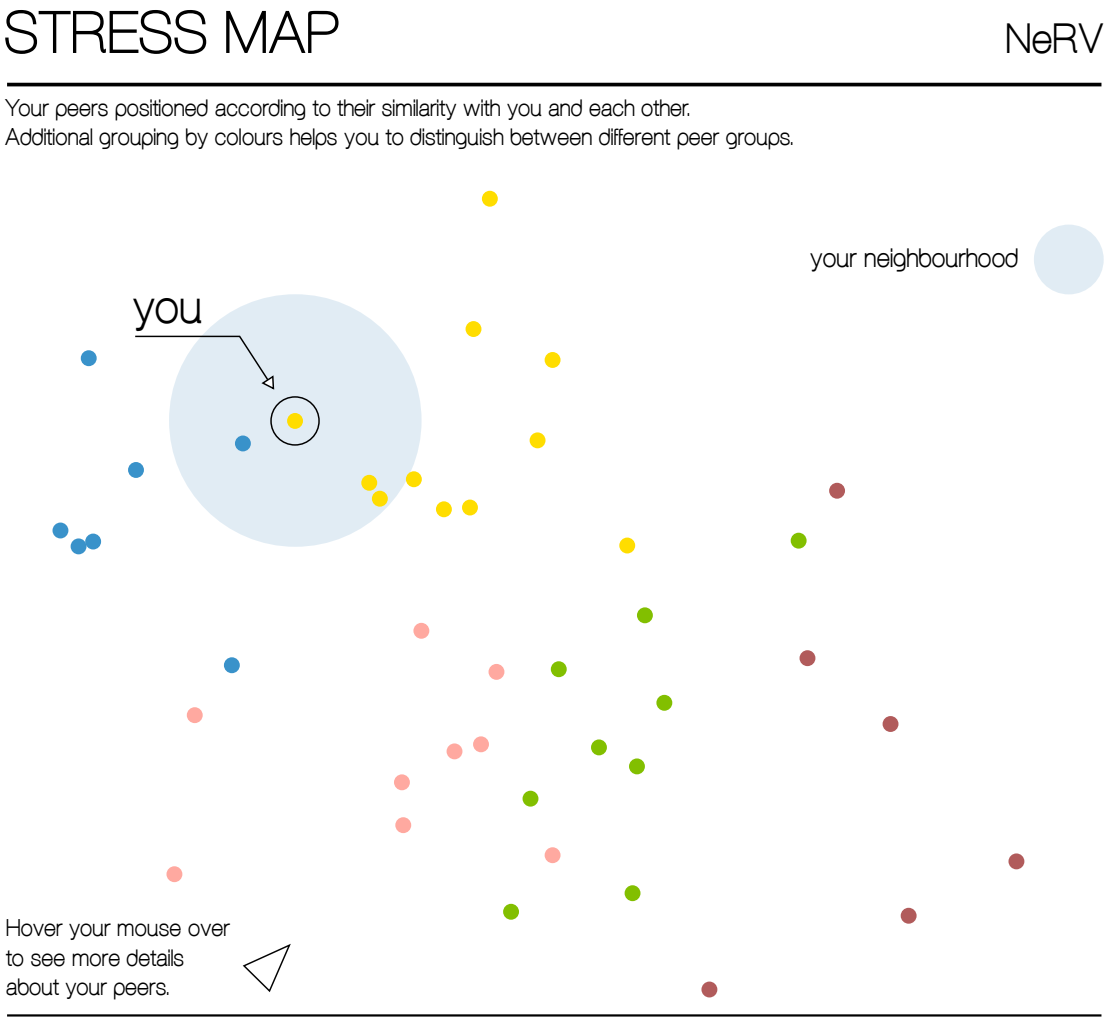


Figure 4.1: The map view of the Stress Map service produced with the NeRV algorithm.

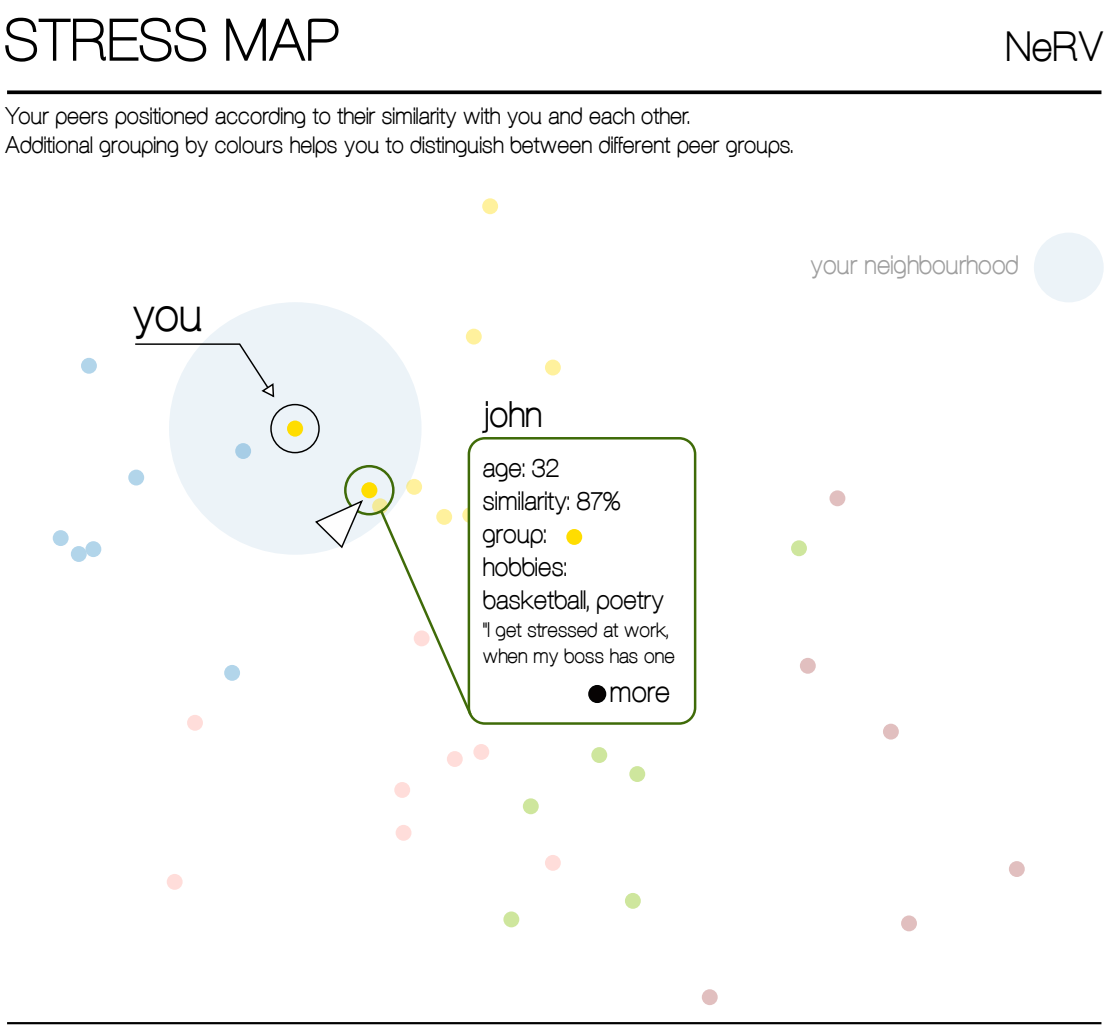


Figure 4.2: The browse view of the Stress Map service produced with the NeRV algorithm.

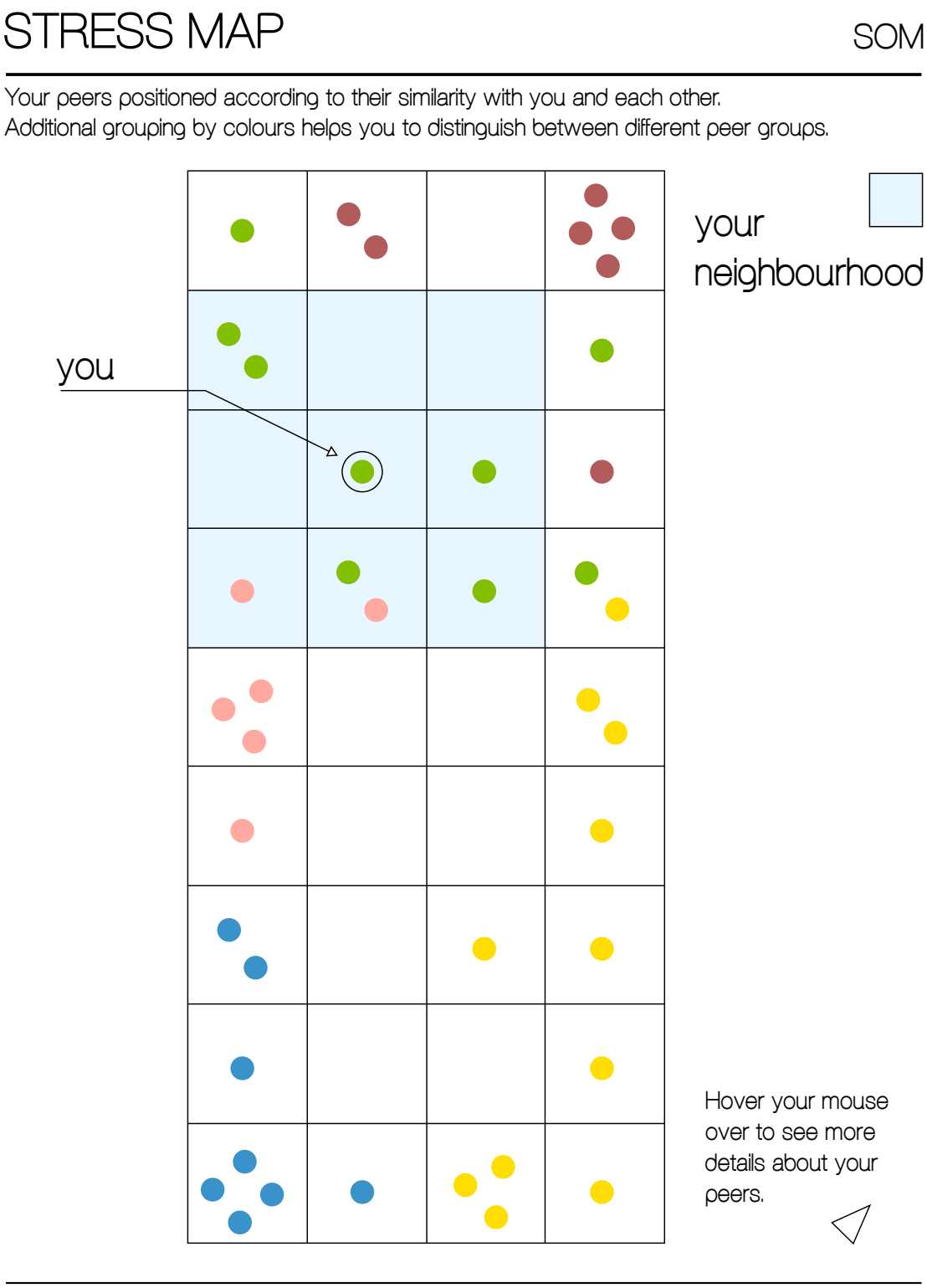


Figure 4.3: The map view of the Stress Map service produced with the SOM algorithm.

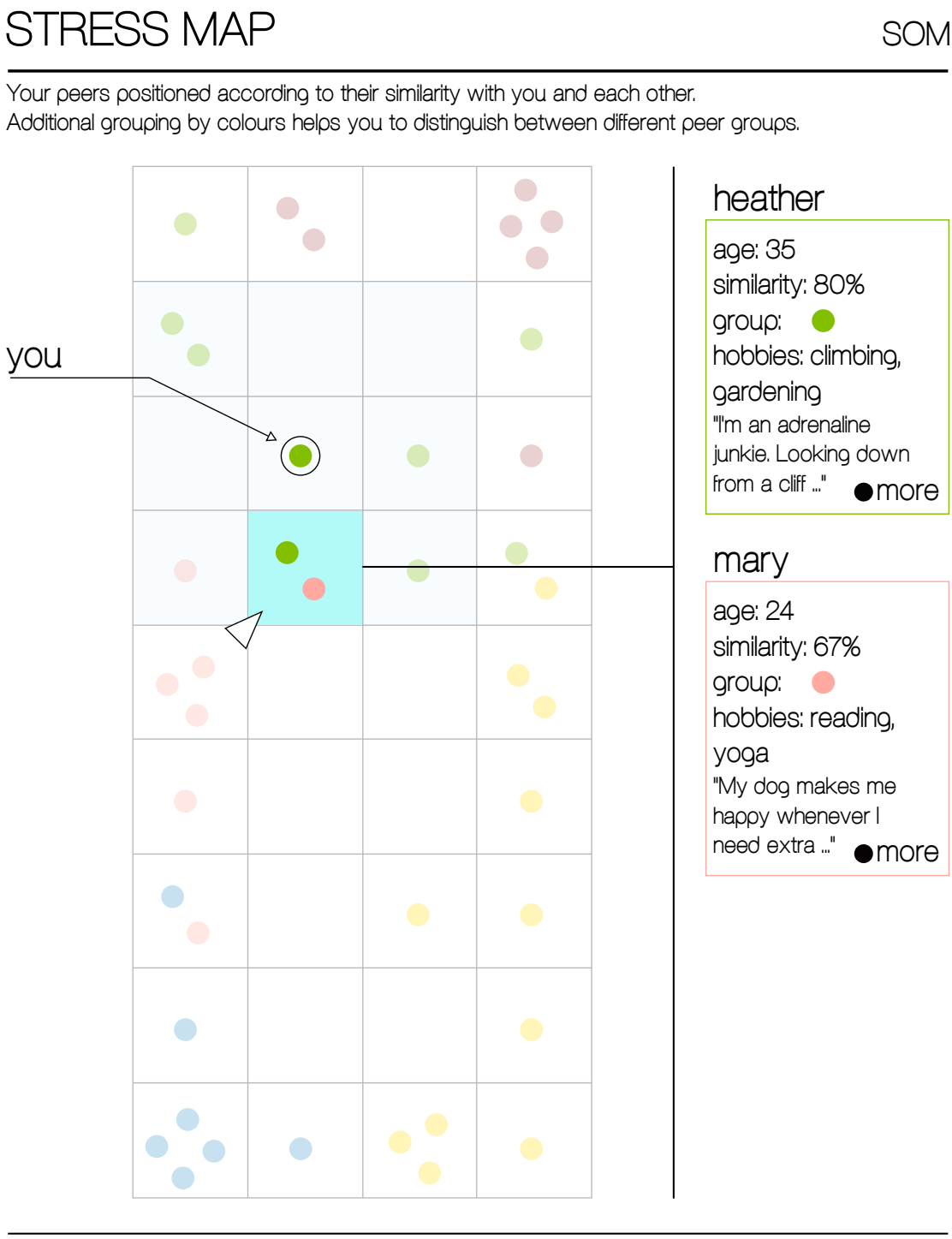


Figure 4.4: The browse view of the Stress Map service produced with the SOM algorithm.

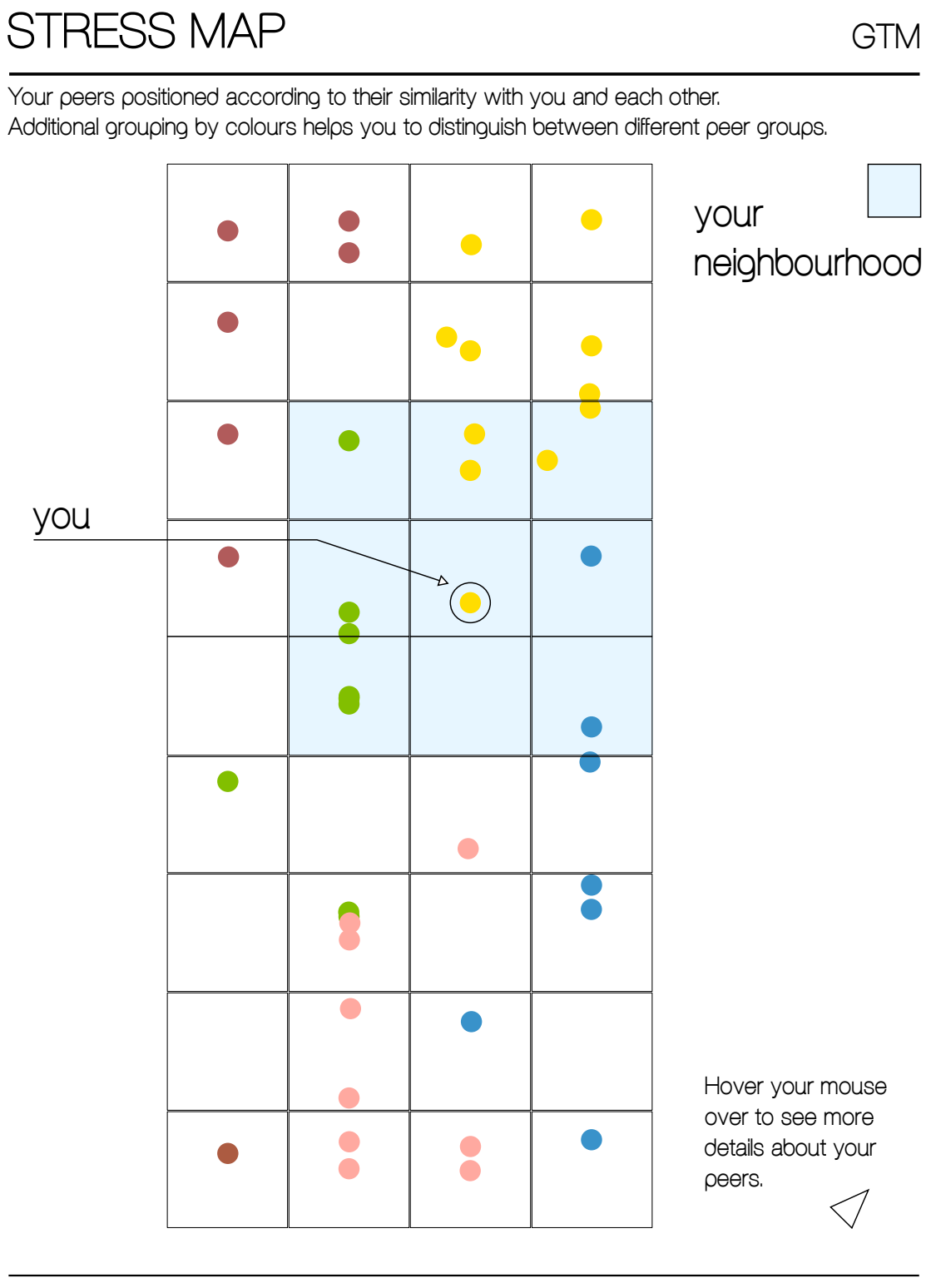


Figure 4.5: The map view of the Stress Map service produced with the GTM algorithm.

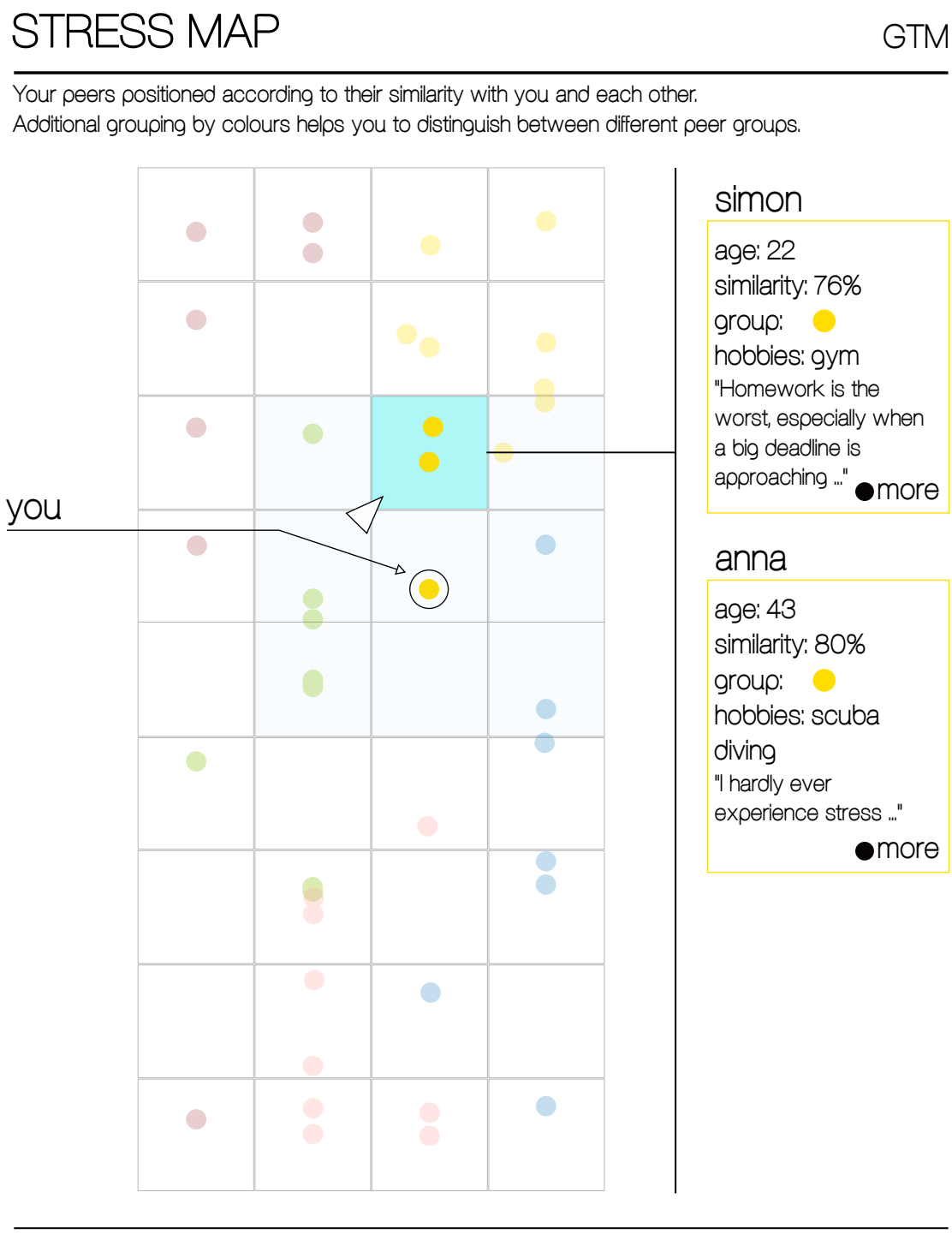


Figure 4.6: The browse view of the Stress Map service produced with the GTM algorithm.

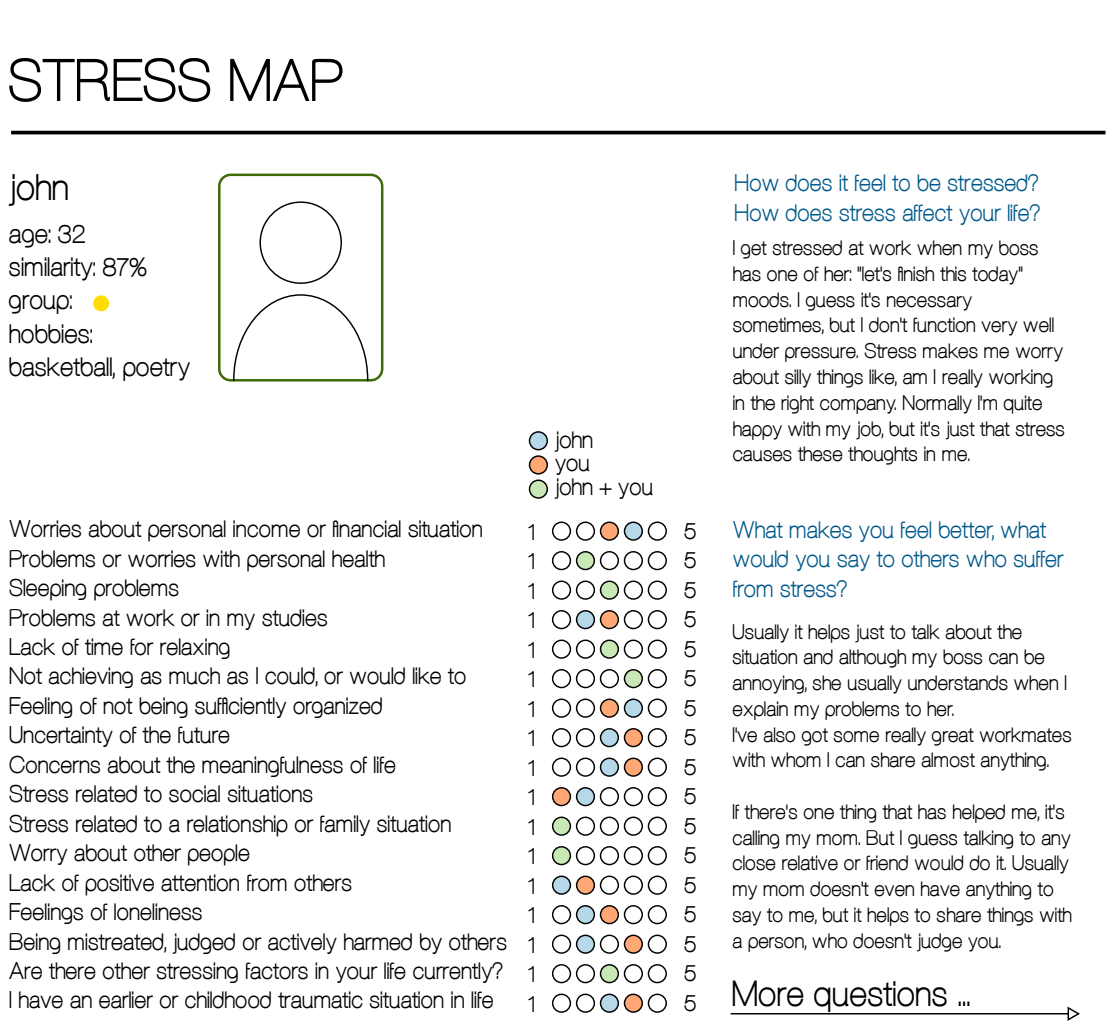


Figure 4.7: An example of a view showing the profile of an artificial user called John.

Chapter 5

User Study

The Stress Map service serves as a mock-up prototype of a real world peer support service. In order to introduce the concept as well as test the ideas with a wider public audience, a small-scale web survey with static images was conducted. The idea was to express the purpose of the service to the audience and then ask them for feedback about the visualisations.

The visualisation methods compared in Chapter 3 produce visualisations that suit different needs. Although a full usability test would have been more informative, the web-based survey was chosen instead, so as to consult first which of the proposed concepts would serve best in a peer support service, before putting more work into a functional prototype.

5.1 The Setup

The study was carried out online as a web form that was created with Google Docs and then embedded on a separate website. The survey was structured so that a short introduction was given first. The visualisations were then added to the website and placed next to the actual web form for convenience, so that it was possible to view the images whilst answering the online questionnaire. A screen capture of the web form is presented in Figure 5.1. The images were grouped such that those corresponding to the NeRV algorithm were shown first, followed by the SOM and finally the GTM. The profile view was pushed to the end, because it is not tied to any of the visualisation algorithms.

The first few questions considered the overall visual representation of the images and were followed by questions regarding usability. Finally some background information was enquired from the respondents. The full set of questions can be found in Appendix B. The questions were partly inspired by the SOM study by Länsiluoto [34].

In order to receive responses quickly, the questionnaire was distributed a number of times via social networks throughout April 2012. A link to the web form was also shared on the wall of the Wellbeing Innovation Camp 2011. In addition, the survey was emailed to the researchers and experts of the VirtualCoach project as well as the Cognitive Complex Systems research group. The questionnaire was first published on April 16th 2012, emails were sent out the day after. The first submissions were received on the same day the form was published. The survey stayed active until the 24th of April.

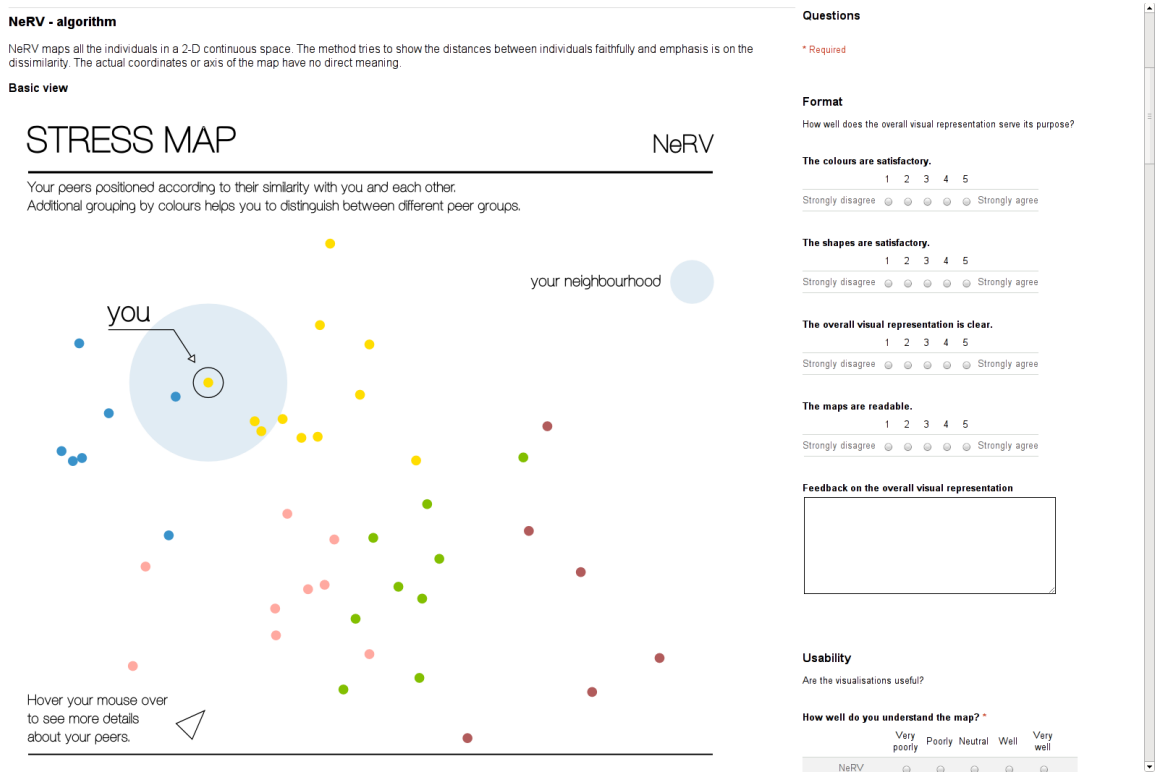


Figure 5.1: A screen capture of the user study website.

5.2 Results

A total of 36 submissions were received, the bulk of which came from acquaintances, friends or family members of the author. Seven respondents identified themselves as members of the VirtualCoach project. Note that the sample turned out somewhat biased in the way the statistics were drawn.

The respondents' ages ranged from 18 to over 50, with the majority being under 30 years old. 23 of the respondents were male, twelve were female and one respondent preferred not to specify their gender. 68% of the respondents stated that their IT skills were 4 or 5 on a scale of 1 (low) to 5 (high). Only a single respondent reported low IT skills. The respondents were mostly engineering, maths or computer science students. Some other occupations or fields of study were reported as well, including research, architecture, entrepreneurship, photography and wellbeing coach. The collated distributions of age, gender and IT skills are shown in Figure 5.2.

Format

The first part of the survey considered the overall visual representation of all the maps. A similar visual appearance was maintained with all the visualisation methods and the purpose of this part of the questionnaire was to assess that. The results of the questions on form are in Figures 5.3–5.6.

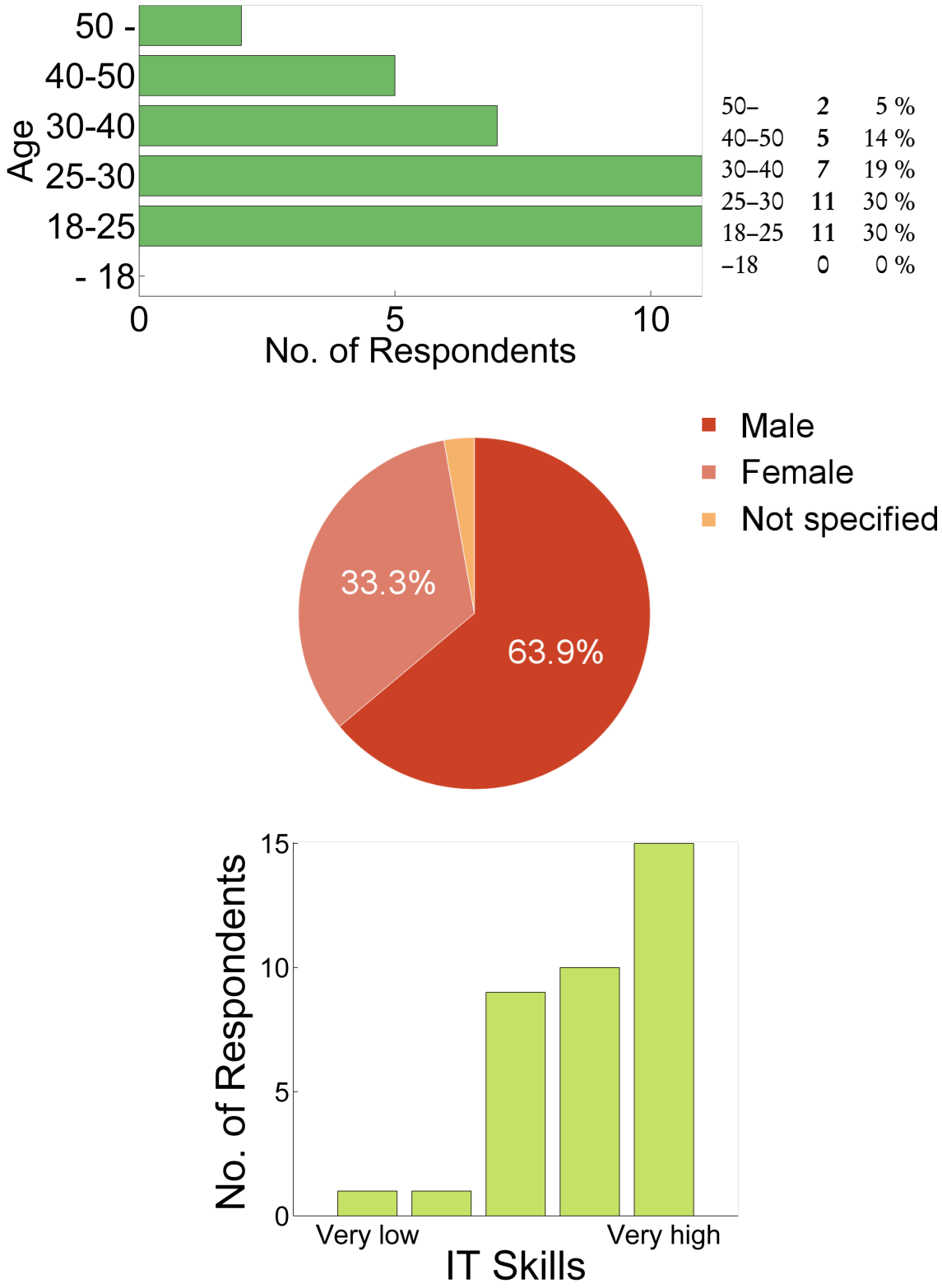
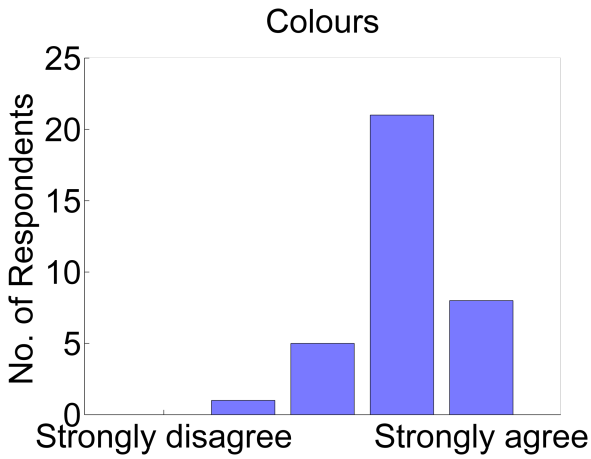


Figure 5.2: The age, gender and reported IT skills of the respondents.

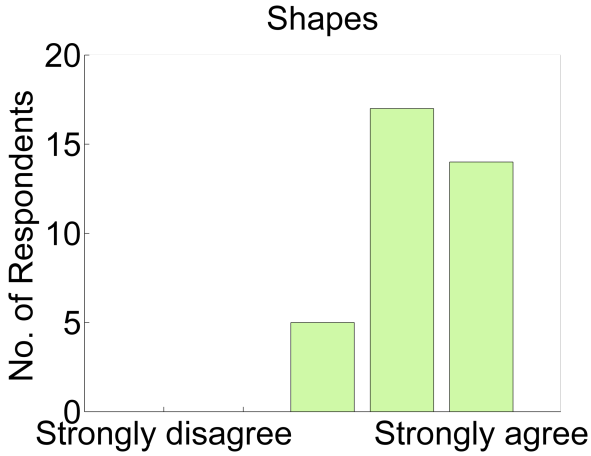
Feedback on the overall visual representation

The respondents were also asked to write comments about the visual appearance of the service. A few responded that the simple appearance was pleasing and that the general



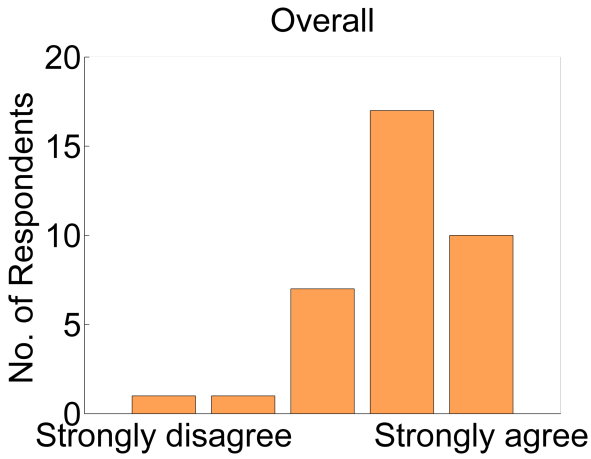
1 – Strongly disagree	0	0 %
2	1	3 %
3	5	14 %
4	21	57 %
5 – Strongly agree	8	22 %

Figure 5.3: The colours are satisfactory.



1 – Strongly disagree	0	0 %
2	0	0 %
3	5	14 %
4	17	46 %
5 – Strongly agree	14	38 %

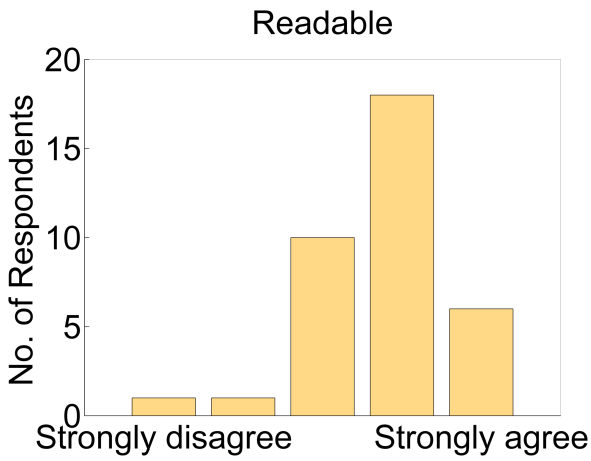
Figure 5.4: The shapes are satisfactory.



1 – Strongly disagree	1	3 %
2	1	3 %
3	7	19 %
4	17	46 %
5 – Strongly agree	10	27 %

Figure 5.5: The overall visual representation is clear.

impression was good. Some were confused about the axes and the fact that they have no meaning. Others were confused about the colouring of the dots and the concept of peer groups in general. One respondent doubted that the chosen colours would be suitable



1 – Strongly disagree	**1**	3 %
2	**1**	3 %
3	**10**	27 %
4	**18**	49 %
5 – Strongly agree	**6**	16 %

Figure 5.6: The maps are readable.

for colour-blind people and suggested a colour scheme with higher contrast. The shapes of the SOM and the GTM map were considered somewhat awkward, and a rectangular map was proposed instead. One respondent wondered whether the peer grouping was based on pre-existing peer groups, such as co-workers and friends, or whether they were based on the answers supplied to the service. A few respondents were concerned that with more data the maps would become more cluttered and hence the grouping more disordered. The overall positioning raised questions as some people found it difficult to judge whether their own stress level was higher or lower compared to their peers.

Usability

The main interest of the survey was to assess the perceived differences between the visualisation methods. The usability part of the questionnaire asked the respondents to evaluate the ease of perceiving aspects like similarity, difference and position of individuals. The results can be found in Figures 5.7, 5.8 and 5.9.

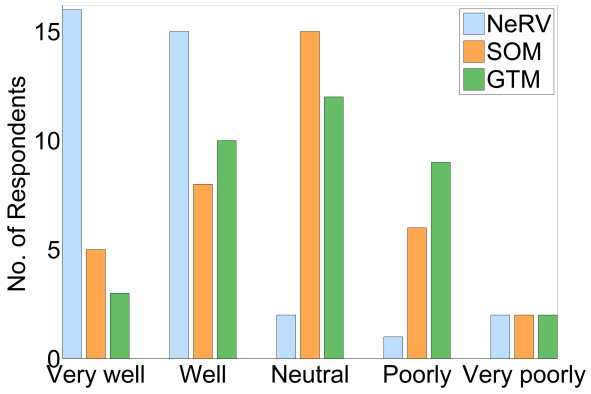
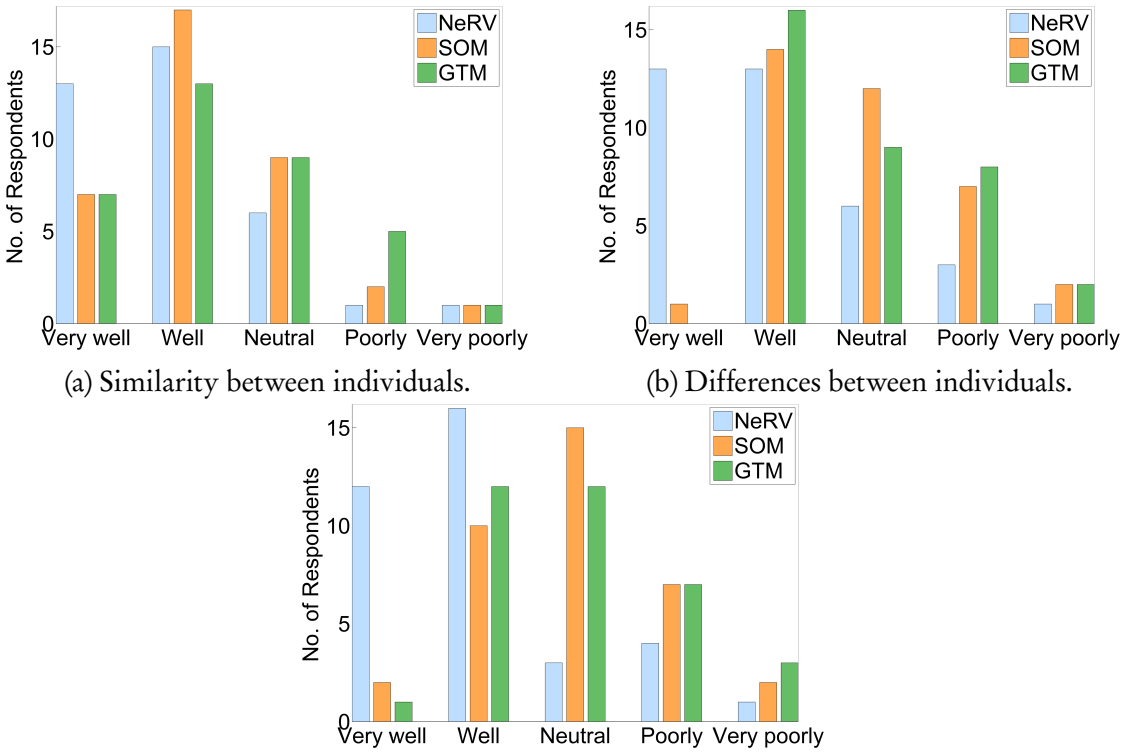


Figure 5.7: How well do you understand the map?



(a) Similarity between individuals.

(b) Differences between individuals.

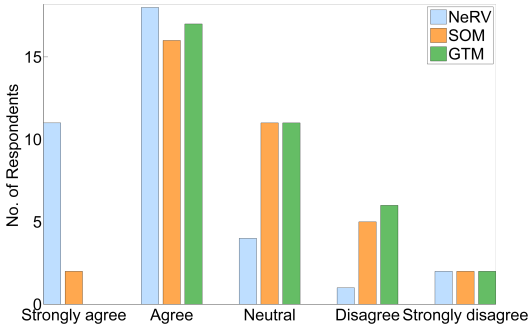
(c) Overall position of individuals.

Figure 5.8: It is easy to perceive and analyse:

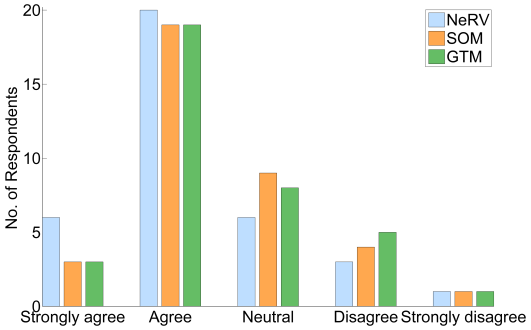
Open comments on the NeRV map

Comments on specific visualisation methods were solicited at the end of the questionnaire. The NeRV map was referred to as "most intuitive", "natural" and was attributed a "continuous feel". According to one response it was easier to understand the similarity of peers in an unlimited space rather than in fixed cells. It was nice and simple to understand, because it did not have any restrictive boxes around. Some argued that the NeRV map was best to present the global structure or for getting an overall picture and that the radial visualisation was more pleasing for the eyes than the cell-based one, as it was more natural.

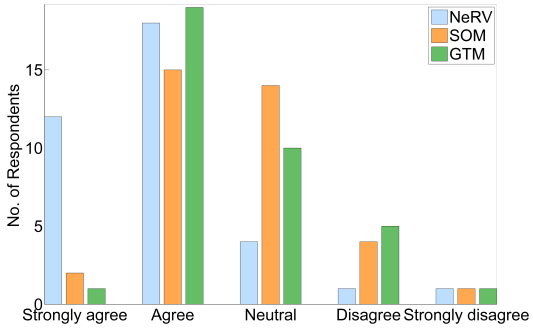
On the negative side, someone felt it was difficult to understand similarities and why some people are closer together. Axes labels as well as further explanation on how to interpret the position of the points was requested in several comments. Some suggested that the texts should be clearer and the peer groups should be explained in the graph, too. One respondent stated that the NeRV map was pleasantly open, but it was overly simple and gave too little information. It was pointed out that both the shape and size of the symbols could be used to express the values of the variables in order to convey more information. Several mouse clicks are required in the current implementation before such information is available.



(a) The map could be used in a peer support service.



(b) The map gives you interesting insight about your peers.



(c) The map helps you to perceive your position amond your peers.

Figure 5.9: General questions about the maps.

Open comments on the SOM map

The SOM inspired further comments. It was described as engineer-like, but readable. The respondents argued that the SOM provided more discrete categorisation, but it was harder to see the bigger picture. It also provided a better concept of neighbourhood and gave a feeling for which characteristics separate one from the nearby crowd. Someone commented that while it was still hard to grasp the concept of similarity, the SOM was somehow still understandable by giving a sense of categorisation. On the other hand, the squares or boxes caused some confusion for some respondents as they were unable to understand the meaning.

Distances between and within cells gave rise to some confusion. One respondent pointed out that it is not clear if the biggest difference is between the far corners of the map or somewhere else. Distances across the map were untrustworthy. The dots inside the cells were not always equidistantly located which confused some people. Once more the meaning of the groups as well as the colours were questioned since there were dots of different colours inside the same cell. The square grid does not clarify whether diagonally adjacent cells are more different than horizontally or vertically adjacent cells. One respondent disliked how the information of everyone in the cell was shown at the same time. As the shape of the grid was unpleasant for some respondents, a version with hexagonal and smaller cells was proposed instead.

Someone argued that the SOM was not meant for consumers or ordinary users. The shape and symbols could again be used to convey more information on the values of the

variables. One respondent requested additional visual explanations as the discretisation of the space might be unclear. It was also stated that the SOM would contain too much information, if the goal was to just see your relative positioning in a group. Similarly to NeRV, the SOM does not explain what each axis represents. In one submission, it was suggested that the categorisation of the SOM should be combined with the more open structure of NeRV to achieve an altogether better visualisation system.

Open comments on the GTM map

The respondents had very similar experiences with the GTM map as they had with the SOM. The grid-like structure and the high and narrow shape of the map were the cause of some confusion. A few respondents remarked that the GTM map was even more confusing than the SOM. One respondent wrote that the GTM map looked like a random hybrid of NeRV and SOM, loosing all the useful features of both. Some points seemed to be clustered very closely together, so that they were difficult to distinguish. This raised again questions concerning the meaning of the cells as well as the colouring. Some of the respondents did not understand why there are cells in the first place if the points can move over the cell boundaries, although somebody concluded that it was more natural to think that people cannot be categorised with clear boundaries. The sliding scale did not seem to add any benefit according to another respondent. In general, the sense of categorisation and still being able to move between categories divided the comments strictly. While some thought it was more natural, others just considered it unpleasant for the eye.

Several points seemed to overlap which was difficult to interpret. One respondent wondered why some points overlapped even though they had different colours and whether the overlapping points represented the maximum available support. Someone asked why the dots were discretised into cells, but then treated differently within each square. Others approved of the sliding boundaries as it illustrated well that people belong to several groups at the same time. Referring to coaching, the map opened up possibilities to find new instructions and peer support outside the common circles, but still close enough to initiate change, it was pointed out by one respondent. Similarly to the SOM, the apparent information overload of the GTM was criticised. Someone stated there is too much mental overhead in wondering how the groups have been formed.

In a general comment, one respondent challenged the idea of quantifying and measuring the concept of stress by arguing that stress is a highly personal and subjective matter. They questioned the legitimacy of the measurement-based process and regarded the whole pursuit to be meaningless or at least ill-posed.

5.3 Discussion

Judging by the results of the multiple-choice questions, it can be said that the visual appearance of the Stress Map was generally satisfying. The majority found the colours and shapes to be satisfactory, although the shape of the SOM and the GTM maps was criticised by some respondents. Clearly the high and narrow grid was an unpopular decision. After the optimal size of the map was determined to be 36 units, the SOM algorithm was left to decide freely the structure and it came up with a 9-by-4 grid. It was decided to maintain the square cells, resulting in an elongated map, which in turn was generally

perceived as unpleasant. 73% of the respondents (strongly) agreed that the overall visual representation was clear, whereas one respondent disagreed and another expressed their strong disagreement. The given answers were somewhat affected by the difficulties experienced in interpreting the maps, with the cell-based maps being most difficult to understand. While 64% considered the maps readable, one person disagreed and one disagreed strongly, 27% had a neutral opinion towards this subject.

The NeRV was generally well understood by the users according to their own opinions. Clearly the SOM and the GTM were more difficult to understand as can be seen from Figure 5.7.

When asked about perceived aspects such as similarity, difference and overall position of individuals, the NeRV was consistently considered to perform better at these tasks. In general, the SOM seems to have performed slightly better than the GTM, although the difference between the two is somewhat negligible. When asked whether a map could be used in a peer support service, the respondents found the NeRV to be the most suitable, whereas the SOM and the GTM were considered significantly less apt. Figure 5.9b shows that nearly all the maps provided the respondents with an equal amount of insight about their peers, so that the corresponding question failed to capture any distinguishing features. The final question was about perceiving one's own position among one's peers. The NeRV was considered to help the most when trying to assess one's own position, while the SOM and the GTM performed a bit worse.

The comments reveal more about the problem of conceptualising and visualising similarity. The most problematic feature for the respondents seemed to be the rectangular grid in the SOM and the GTM, whereas the lack of boxes seemed to be the main advantage of the NeRV. Admittedly the use of boxes was perhaps more of a design decision rather than an inherent requirement of the methods, although the presence of a small number of units onto which the data points are mapped is indeed a characteristic of both the SOM and the GTM. However, the maps could have been visualised in a different way to make them look more attractive. Hexagonal cells were suggested and indeed the SOM has been reported to perform better with a hexagonal grid, which also lacks the ambiguity between vertical, horizontal and diagonal neighbours. However, a common user might find the hexagonal grid odd and artificial and thus a more familiar looking rectangular grid was chosen. As the service gathers more users and more data, the map size should be increased accordingly, so as to have more map units, and thus the map becomes more granular and less discrete. Still there is no escaping the fact that the SOM and the GTM are more discrete methods by construction. They rely on a set of prototype units, which represent the original data in the visualisation space, whereas the NeRV maps the data points directly onto the visualisation space.

Unlabelled axes were the cause for some confusion, even though it was explained in the questionnaire that the methods are nonlinear and thus the directions in the resulting visualisations do not have any straightforward meaning. The purpose of the nonlinear methods is to retrieve and visualise the relevant neighbours of each individual in the best possible fashion. Emphasis is given to the local distances as well as relationships between neighbouring points. Perhaps this should be explained more clearly or be more evident in the visualisation. One way to overcome this limitation would be to label some of the areas in the map, so that users could judge their own positions relative to these labelled areas (see, e.g., [30]). However, labelling and grouping the data is a non-trivial task. There are a

number of ways to accomplish this unsupervised machine learning task, but the analysis of clustering or labelling algorithms is beyond the scope of this thesis.

A tentative grouping was performed using the very simple k-means clustering algorithm. The Stress Data was clustered into five groups based on the Euclidean distance between each of the data points. The group membership was visualised with colours so that the users could have a sense of specific areas or groups in the maps. While the colours were meant to visualise a certain grouping, the location on the map was meant to provide a separate way of seeing the relationships. However, some respondents were confused by the fact that some points were mapped into the same cell in the SOM or in the GTM, even though they were attributed to different groups. Recall that one respondent asked whether the groups were some pre-existing peer groups, such as co-workers or friends. This was not the case, but it does suggest a potentially clearer way of grouping. However, the groups might then be scattered around the map, since the members of a certain group could be very different on a different level. This is again more of a design question and should be considered along with other alternatives.

Although the NeRV map was praised for its simplicity, it is fair to say that it contains only little information. Some valuable insight could perhaps be conveyed by varying the symbols somehow without sacrificing too much of the clarity. But then again it was also stated that the SOM and the GTM contain too much information, the argumentation being that the users do not want to think about why each individual is placed in a certain box if they are only interested in their own position with respect to that of their neighbours. Other respondents thought that the categorisation of the SOM and the GTM was good and that it somehow helped to understand and perceive their own neighbourhood.

The collated results for the multiple-choice questions and the corresponding comments show that the visualisation made with the NeRV algorithm was judged to be the most adequate one. As mentioned above, this success cannot be ascribed directly to the NeRV as the MDS would have resulted in a very similar visualisation. The only difference would have been the ordering and positioning of the individuals. The users were not able to explore the map freely and so they could not judge whether some ordering was better than another. The winner of this survey was the continuous-projection-type visualisation rather than a particular algorithm itself. In addition, it can be argued that with a few modifications the visualisations of the SOM and the GTM could be made more attractive. By developing the visualisations based on the comments from the presented survey, the methods could perhaps perform better against the NeRV and the MDS, and so further investigation is encouraged.

Suggestions

- Simple is better.
- The peer groups could be coloured according to existing groupings (colleagues, friends, ...).
- The colouring scheme should make sense for people with colour blindness. This can be achieved, e.g., using strong contrasts.
- Using hexagonal cells should be considered when using the SOM and the GTM. The cell boundaries could be left out completely.

- If using a grid, the aspect ratio should be set closer to the golden ratio or so that the image becomes more rectangular.
- More information such as the values of the variables could be conveyed using the shape and size of the symbols.
- The lack of labelled axes could be compensated by labelling areas of the map to help the interpretation.
- Experts should be consulted when selecting features that represent the data. If the experts do not have extensive experience in data analysis initial visualisations should be used when explaining the purpose of the analysis to the experts.

The survey was successful in figuring out which of the presented methods would serve best in a peer support service. It also emerged that the concept of a peer support service is rather difficult one to understand as well as and explain. The users need to realise the benefits of using such a service before they will invest time to it. In this survey, the users had difficulties in understanding the visualisation that was provided for finding peers. The whole concept of an online peer support service with automated features is still under development. The comments gathered in this survey serve as an excellent starting point when considering further developments. The users were asked to imagine the use of a peer support service in this study, which probably added to the confusion as the concept of online peer support is not yet familiar to everyone. A working prototype of such a service would be essential to gain the full benefit of an extended user study. However, conducting a user study, even one with static images, was necessary to examine concretely the idea of visualising peers.

Chapter 6

Conclusions

Early detection and prevention of health problems, and services that support the spontaneous empowerment of the individual, are key issues in the future of healthcare. Joining the expertise of the statistical analysis community, wellbeing professionals and aspiring entrepreneurs is the mission of the VirtualCoach initiative. One approach of the project has been to develop new kinds of social services that incorporate peer support along with advanced data analysis methods. This thesis was written to support the VirtualCoach project in its progress towards this goal.

The main questions of the thesis were: what does it mean to visualise peers, what methods could be used and what should be taken into account when using them? The first question was discussed in Chapter 2 along with the reasons, why peer support was chosen as a starting point. The second question was answered in Chapter 3, where four dimensionality reduction algorithms were presented and compared. Chapters 4 and 5 try to answer the last question.

this thesis assume a distance based model of similarity.

Peer support and peer stories can have a significant impact on one's wellbeing when facing a life-changing situation. Peer support is known to work but the mechanisms and reasons behind it are not thoroughly agreed upon, although some psychosocial processes have been suggested to explain the beneficial effects. The Internet offers vast possibilities to connect with peers, whilst changing the problem of finding peers into a problem of finding just the right peers. Advanced data analysis methods may assist in the task by locating and evaluating potential peers and providing the user with sophisticated visualisations. The evaluation of peers is based on assessing the similarity of people. Field experts may be consulted in order to find a suitable set of features to represent the objects under study. Similarity as peers is then evaluated by computing a distance between these feature vectors.

It is difficult for humans to understand multidimensional data, that is the complex relations between the variables and the relational structure of the data objects. However, with dimensionality reduction methods the data may be analysed without resorting to only a subset of the available features. Thus all the information in the data can be taken into account and possibly more insight can be gained from the data. Visualising the data is one way to use dimensionality reduction to analyse multidimensional data. Visual representations of complex data are easier to interpret because they reduce the cognitive load of the user by externalising part of the information processing.

Linear dimensionality reduction methods were not considered in this thesis, as the task of ordering individuals based on their similarity is a rather complex one. Instead of finding an interpretable global representation of the locations of the objects, more emphasis was put on retrieving the local neighbourhood structures. The four nonlinear methods presented in this thesis were divided into two categories: the nonlinear projection methods represented by MDS and the NeRV and the topographic map methods represented by the SOM and the GTM. The methods were compared quantitatively using a similar set of goodness measures that were used to develop the NeRV. The NeRV performed better than the MDS in the tests, but there was no clear difference between the performance of the SOM and the GTM. Based on the quantitative analysis, the NeRV seemed more favourable than the MDS for use in visualisation. The SOM and the GTM were both chosen for the next phase as they are able to provide different visualisations. An extensive comparison of different dimensionality reduction methods was out of the scope of this thesis. The purpose of the quantitative analysis was to confirm that the methods could be used in visualisation and that their performance could be tested at least in principle.

The selected methods were then used to produce visualisations for a hypothetical peer support service, that was designed around the visualisations. The service was called the Stress Map and its main purpose was to gather information of the users and then visualise their relative proximity to each other based on this information. The visualisations provided a sort of mock-up prototype of the service, without any actual functionality. The service was designed to show the methods in use and to explicitly express the idea of visualising peers.

In order to answer the final question of what to take into account when visualising peers, a small user survey was conducted. The prototype peer support service was presented to a group of users and feedback on the design and the use of the methods was gathered. The respondents came up with many valuable comments and suggestions regarding the visualisation and the usability of the service. However, many users paid attention solely to the visual aspects of the service and the purpose of the prototype or the survey was not clearly understood. The fact that the concept of a peer support service was not completely self-explanatory to the respondents is an important finding as well. It means the concept and the delivery of the message need to be developed further.

All in all, the NeRV based visualisation was considered most appealing and suitable for presenting peer relationships. Some respondents were concerned about the lack of labelled axes and it was also pointed out that the visualisation contains not enough information. The user interface should be developed so that more relevant information would be accessible with fewer mouse clicks. As suggested, the shape and size of the symbols could be used to visualise the values of the variables of each point. It should be noted that the MDS would produce a very similar visualisation, thus in principle, any of the two would have resulted in the same feedback.

The SOM and the GTM were considered a little awkward and the use of rectangular cells divided strongly the opinions of the respondents. Some liked the sense of categorisation, while others objected to the confined feeling of the cells. A more square shaped grid and the use of hexagonal map units was suggested. An advantage of the SOM and the GTM is that the visualisations need not be computed each time a new data point is inserted into the map, whereas the NeRV and the MDS do not share this property.

Many respondents had difficulties in understanding the colouring of the dots, which

was clearly not explained well enough. Also the very purpose of using automated data analysis methods to determine similarity and the use of this kind of similarity measure was questioned. Quantifying complex phenomena like the ones related to human well-being, is surely not a trivial task and perhaps not possible to achieve with the present methods. However, to be able to develop something new, it is necessary to first formulate the ideas and even present them with the help of visualisations. Basing the discussion on something tangible is a better starting point than not having anything to look at in the first place.

Some initial plans to integrate the NeRV visualisation to the prototype designed in the VirtualCoach project have already been made [23]. Perhaps in the future such visualisations will be used to cleverly present multidimensional data in different services. Hopefully mapping complex relationships like the similarity between people will be commonplace and systems that enable people to find help from just the right place are available to all.

Appendix A

Stress Questionnaire

The Stress questionnaire was designed by Lagus et al. [33] and was conducted at the WIC-2011 Wellbeing Innovation Camp [4] in 2011 as part of the VirtualCoach Project. The 43 respondents are students, who participated in the WIC-2011. The survey consists of 33 multiple-choice questions and 5 open text questions. Answers to the questions were given on a Likert-scale from 1 to 5.

A.1 Multiple-choice questions

When you think about your life for the past month, how often did you experience the following?

- Not at all (1), Rarely (2), Sometimes (3), Often (4), All the time (5).

1. Worries about personal income or financial situation
2. Problems or worries with personal health
3. Sleeping problems
4. Problems at work or in my studies
5. Lack of time for relaxing
6. Not achieving as much as I could, or would like to
7. Feeling of not being sufficiently organized
8. Uncertainty of the future
9. Concerns about the meaningfulness of life
10. Stress related to social situations
11. Stress related to a relationship or family situation
12. Worry about other people

13. Lack of positive attention from others
14. Feelings of loneliness
15. Being mistreated, judged or actively harmed by others
16. Are there other stressing factors in your life currently?
17. I have an earlier or childhood traumatic situation in life

When you think about your life for the past month, how often do you use the following stress relieving methods?

- Not at all (1), Rarely (2), Sometimes (3), Often (4), All the time (5).

1. Reducing activities
2. Resting or sleeping more
3. Physical exercising
4. Nature
5. Relaxation or mindfulness techniques
6. Eating more
7. Eating more healthily
8. Alcohol, cigarettes or drugs
9. Shopping
10. Spending time alone
11. Music, literature or art
12. TV, Internet or videogames
13. Trying to eradicate the cause of stress
14. Mental activity about the cause of stress (brainstorming, problem solving, planning etc.)
15. Leaving the stressful environment or situation
16. Accepting things the way they are
17. Letting my feelings and emotions out
18. Humor
19. Spending time with family or friends

20. Massage, hugging or touching others
21. Sexual activities
22. Religion or spiritual practice
23. Counseling
24. Helping others

Other means of stress relieving methods was asked with an open question: other means, please list.

A.2 Open text questions

1. Do you remember a time when you have been very stressed? Please describe the situation.
2. How does it feel to be stressed? How does stress affect your life?
3. When you were very stressed and were then able to relax, what happened? What created the change for the better?
4. Imagine a life where you experience occasional stress, but do not suffer from it. What is your life like?
5. What would you say to others who suffer from stress?

Appendix B

Stress Map User Study

In order to introduce the concept as well as test the ideas with a wider public audience, a small-scale web survey with static images was conducted. The study was carried out online as a web form that was created with Google Docs and then embedded on a separate website. The survey was structured so that a short introduction was given first. The visualisations were then added to the website and placed next to the actual web form for convenience, so that it was possible to view the images whilst answering the online questionnaire. The questions are listed in this Appendix.

B.1 Format

How well does the overall visual representation serve its purpose?

Rate from 1 (Strongly agree) to 5 (Strongly disagree).

1. The colours are satisfactory.
2. The shapes are satisfactory.
3. The overall visual representation is clear.
4. The maps are readable.

Feedback on the overall visual representation.

B.2 Usability

How well do you understand the map?

	Very poorly	Poorly	Neutral	Well	Very well
NeRV	o	o	o	o	o
SOM	o	o	o	o	o
GTM	o	o	o	o	o

Looking at the NeRV map, it is easy to perceive and analyse:

	Very poorly	Poorly	Neutral	Well	Very well
Similarity between individuals	o	o	o	o	o
Differences between individuals	o	o	o	o	o
Overall position of individuals	o	o	o	o	o

Looking at the SOM map, it is easy to perceive and analyse:

	Very poorly	Poorly	Neutral	Well	Very well
Similarity between individuals	o	o	o	o	o
Differences between individuals	o	o	o	o	o
Overall position of individuals	o	o	o	o	o

Looking at the GTM map, it is easy to perceive and analyse:

	Very poorly	Poorly	Neutral	Well	Very well
Similarity between individuals	o	o	o	o	o
Differences between individuals	o	o	o	o	o
Overall position of individuals	o	o	o	o	o

The map could be used in a peer support service.

	Very poorly	Poorly	Neutral	Well	Very well
NeRV	o	o	o	o	o
SOM	o	o	o	o	o
GTM	o	o	o	o	o

The map gives interesting insight about your peers.

	Very poorly	Poorly	Neutral	Well	Very well
NeRV	o	o	o	o	o
SOM	o	o	o	o	o
GTM	o	o	o	o	o

The map helps you to perceive your position among your peers.

	Very poorly	Poorly	Neutral	Well	Very well
NeRV	o	o	o	o	o
SOM	o	o	o	o	o
GTM	o	o	o	o	o

Open comments on the NeRV map.

Open comments on the SOM map.

Open comments on the GTM map.

B.3 Background information

Age

- – 18
- 18 – 25
- 25 – 30
- 30 – 40
- 40 – 50
- 50 –

Gender

- Not specified
- Male
- Female

What is your occupation or field of studies?

Rate your IT skills from 1 to 5.

- Very low 1...5 Very high

Are you a member of the VirtualCoach project?

Bibliography

[1] dredviz: dimensionality reduction for information visualization. http://research.ics.tkk.fi/mi/software/dredviz/. [Online; accessed 9-May-2012].

[2] Netlab neural network software. http://www1.aston.ac.uk/eas/research/groups/ncrg/resources/netlab/. [Online; accessed 27-June-2012].

[3] Virtualcoach project website. http://blog.pathsofwellbeing.com/. [Online; accessed 26-July-2012].

[4] Wellbeing Innovation Camp. http://www.cis.hut.fi/wicamp/. [Online; accessed 20-August-2012].

[5] Mikko Berg. *Human Abilities to Perceive, Understand, and Manage Multi-Dimensional Information with Visualizations*. PhD thesis, Aalto University, 2012.

[6] L F Berkman. The role of social relations in health promotion. *Psychosomatic Medicine*, 57(3):245–254, 1995. URL http://www.ncbi.nlm.nih.gov/pubmed/7652125.

[7] Christopher M Bishop, Markus Svensén, and Christopher K I Williams. GTM: The Generative Topographic Mapping. *Neural Computation*, 10(1):215–234, 1998. URL http://eprints.aston.ac.uk/1128/.

[8] Ingwer Borg and Patrick J. F. Groenen. *Modern Multidimensional Scaling: Theory and Applications (Springer Series in Statistics)*. Springer, second edition, August 2005. ISBN 0387251502. URL http://www.worldcat.org/isbn/0387251502.

[9] H.Sharon Campbell, Marie Rose Phaneuf, and Karen Deane. Cancer peer support programs—do they work? *Patient Education and Counseling*, 55(1):3 – 15, 2004. ISSN 0738-3991. doi: 10.1016/j.pec.2003.10.001. URL http://www.sciencedirect.com/science/article/pii/S073839910300301X.

[10] Daniel M. Ennis, F. Gregory Ashby. Similarity measures. *Scholarpedia*, 2(12):4116, 2007.

[11] A. Frank and A. Asuncion. UCI machine learning repository, 2010. URL http://archive.ics.uci.edu/ml.

[12] John Gantz and David Reinsel. The 2011 digital universe study: Extracting value from chaos. http://www.emc.com/collateral/demos/microsites/emc-digital-universe-2011/index.htm. [Online; accessed 23-April-2012].

[13] A. J. Gartner and F. Riessman. Self-help and mental health. *Hospital & community psychiatry*, 33(8):631–635, aug 1982. ISSN 0022-1597. URL http://view.ncbi.nlm.nih.gov/pubmed/7118097.

[14] Robert Goldstone. Similarity. In Robert A Wilson and Frank C Keil, editors, *The MIT Encyclopedia of the Cognitive Sciences*, pages 763–765. MIT Press, 2001. URL http://www.mitpressjournals.org/doi/pdfplus/10.1162/coli.2000.26.3.463.

[15] Patrick J. F. Groenen and Michel van de Velden. *Multidimensional Scaling*. John Wiley & Sons, Ltd, 2005. ISBN 9780470013199. doi: 10.1002/0470013192.bsa415. URL http://dx.doi.org/10.1002/0470013192.bsa415.

[16] Geoffrey Hinton and Sam Roweis. Stochastic neighbor embedding. *Advances in neural information processing systems*, 15:833–840, 2003. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.13.7959&rep=rep1&type=pdf.

[17] Timo Honkela, Jorma Laaksonen, Hannele Törrö, and Juhani Tenhunen. Media map: A multilingual document map with a design interface. In *Advances in Self-Organizing Maps - Proceedings of WSOM 2011, 8th International Workshop*, pages 247–256, 2011.

[18] H Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417–441, 1933. URL http://content.apa.org/journals/edu/24/6/417.

[19] J S House, D Umberson, and K R Landis. Structures and processes of social support. *Annual Review of Sociology*, 14(1):293–318, 1988. URL http://arjournals.annualreviews.org/doi/abs/10.1146/annurev.so.14.080188.001453.

[20] Samuel Kaski. Data exploration using self-organizing maps. *Acta Polytechnica Scandinavica, Mathematics, Computing and Management in Engineering Series No. 82*, March 1997. D.Sc.(Tech.) Thesis, Helsinki University of Technology, Finland.

[21] Samuel Kaski and Krista Lagus. Comparing self-organizing maps. In Christoph von der Malsburg, Werner von Seelen, Jan Vorbrüggen, and Bernhard Sendhoff, editors, *Artificial Neural Networks — ICANN 96*, volume 1112 of *Lecture Notes in Computer Science*, pages 809–814. Springer Berlin / Heidelberg, 1996. ISBN 978-3-540-61510-1. URL http://dx.doi.org/10.1007/3-540-61510-5_136.

[22] Samuel Kaski, Timo Honkela, Krista Lagus, and Teuvo Kohonen. WEBSOM—self-organizing maps of document collections. *Neurocomputing*, 21: 101–117, 1998.

[23] Jussa Klapuri, Lassi Haaranen, and Ilari T. Nieminen. Questionnaire prototype designed at the VirtualCoach project. 2012. URL http://pathsofwellbeing.com.

[24] T. Kohonen, J. Hynninen, J. Kangas, and J. Laaksonen. SOM PAK: The Self-Organizing Map program package, 1996. URL http://www.cis.hut.fi/research/som_lvq_pak.shtml.

[25] Teuvo Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59–69, 1982.

[26] Teuvo Kohonen and Panu Somervuo. How to make large self-organizing maps for nonvectorial data. *Neural Networks*, 15(8–9):945 – 952, 2002. ISSN 0893-6080. doi: 10.1016/S0893-6080(02)00069-2. URL http://www.sciencedirect.com/science/article/pii/S0893608002000692.

[27] J. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29:1–27, 1964. ISSN 0033-3123. URL http://dx.doi.org/10.1007/BF02289565.

[28] J. Kruskal. Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29:115–129, 1964. ISSN 0033-3123. URL http://dx.doi.org/10.1007/BF02289694.

[29] Krista Lagus. Text mining with the WEBSOM. *Acta Polytechnica Scandinavica, Mathematics and Computing Series No. 110*, December 2000. D.Sc.(Tech) Thesis, Helsinki University of Technology, Finland.

[30] Krista Lagus and Samuel Kaski. Keyword selection method for characterizing text document maps. In *Proceedings of ICANN99, Ninth International Conference on Artificial Neural Networks*, volume 1, pages 371–376. IEE, London, 1999.

[31] Krista Lagus and Juho Saari. Peer support in the era of internet and social media. Personal communication, draft for an article, jul 2012.

[32] Krista Lagus, Timo Honkela, Samuel Kaski, and Teuvo Kohonen. WEBSOM for textual data mining. *Artificial Intelligence Review*, 13(5/6):345–364, December 1999.

[33] Krista Lagus, Tuula Styrman, and Zaur Izzatdust. Stress and relaxation questionnaire. unpublished, 2011.

[34] Aapo Länsiluoto. *Economic and Competitive Environment Analysis in the Formulation of Strategy*. PhD thesis, Turku School of Economics and Business Administration, 2004.

[35] S Mead, D Hilton, and L Curtis. Peer support: a theoretical perspective. *Psychiatric Rehabilitation Journal*, 25(2):134–141, 2001. URL http://www.ncbi.nlm.nih.gov/pubmed/11769979.

[36] Douglas L Medin, Robert L Goldstone, and Dedre Gentner. Respects for similarity. *Psychological Review*, 100(2):254–278, 1993. URL http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-295X.100.2.254.

[37] Yasir Mehmood, Mudassar Abbas, Xi Chen, and Timo Honkela. Self-organizing maps of nutrition, lifestyle and health situation in the world. In *Advances in Self-Organizing Maps - Proceedings of WSOM 2011, 8th International Workshop*, pages 160–167. Springer, 2011.

[38] Catherine Panzarella, Lauren B Alloy, and Wayne G Whitehouse. Expanded hopelessness theory of depression : On the mechanisms by which social support protects against depression. *Cognitive Therapy and Research*, 30(3):307–333, 2006. URL http://www.springerlink.com/index/10.1007/s10608-006-9048-3.

[39] Roger Shepard. The analysis of proximities: Multidimensional scaling with an unknown distance function. i. *Psychometrika*, 27:125–140, 1962. ISSN 0033-3123. URL http://dx.doi.org/10.1007/BF02289630.

[40] P. Solomon. Peer support/peer provided services underlying processes, benefits, and critical ingredients. *Psychiatr Rehabil J*, 27(4):392–401, 2004. ISSN 1095-158X. URL http://view.ncbi.nlm.nih.gov/pubmed/15222150.

[41] P A Thoits. Stress, coping, and social support processes: where are we? what next? *Journal of Health and Social Behavior*, Spec No(1995):53–79, 1995. URL http://www.ncbi.nlm.nih.gov/pubmed/7560850.

[42] Warren S. Torgerson. *Theory and Methods of Scaling*. John Wiley & Sons, 1958. ISBN 0471879452. URL http://www.worldcat.org/isbn/0471879452.

[43] Edward R. Tufte. *The visual display of quantitative information*. Graphics Press, Cheshire, CT, USA, 1986. ISBN 0-9613921-0-X.

[44] Amos Tversky. Features of similarity. *Psychological Review*, 84(4):327–352, 1977. URL http://psycnet.apa.org/journals/rev/84/4/327/.

[45] A L Vangelisti. Challenges in conceptualizing social support. *Journal of Social and Personal Relationships*, 26(1):39–51, 2009. URL http://spr.sagepub.com/cgi/doi/10.1177/0265407509105520.

[46] Jarkko Venna and Samuel Kaski. *Nonlinear dimensionality reduction as information retrieval*, volume 7, page 568–575. Omnipress, 2007. URL http://eprints.pascal-network.org/archive/00003560/.

[47] Jarkko Venna, Jaakko Peltonen, Kristian Nybo, Helena Aidos, and Samuel Kaski. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *J. Mach. Learn. Res.*, 11:451–490, March 2010. ISSN 1532-4435.

[48] Juha Vesanto, Johan Himberg, Esa Alhoniemi, and Juha Parhankangas. Self-organizing map in Matlab: the SOM toolbox. In *In Proceedings of the Matlab DSP Conference*, pages 35–40, 1999.

[49] World Design Capital Helsinki. Visualized information for residents. http://wdchelsinki2012.fi/en/news/2012-02-20/visualized-information-residents, 2012. [Online; accessed 23-April-2012].