VISUALIZATION OF LARGE TEMPORAL SOCIAL NETWORK DATASETS

Uraz Cengiz TURKER

CS, Ms Thesis, 2008

Thesis Supervisor: Assistant Professor Selim Saffet BALCISOY

Keywords: Information Visualization, Hyperbolic Layout, Social Networks, Anomaly Detection.

Abstract

Social network datasets consist of what sociologists call 'social structures', accumulation of all communication channels that social actors share ideas and information between each other. Social network analysis reveals characteristics and properties of social networks by applying specific metrics. Although, size of a real-life social network dataset can reach millions of relations belong to millions of social actors with large temporal dimension, existing information visualization tools can represent at most several thousands of these actors.

This thesis presents a conceptual design study focused on visualization of large temporal social network datasets with a novel visualization method. Proposed technique combines *Ideal Gas Law* (IGL) with Jacob Moreno's theory of *The Cannon of Creativity* to layout social network datasets in 3D hyperbolic space and can render 50,000 social actors at interactive speed. A proof-of-the-concept program is developed around this technique allowing users to perform several analysis tasks on temporal social network datasets. Users can explore the network, control the amount of visual clutter, and identify communication anomalies in run time. Moreover, they can search a specific actor and visually follow her communication pattern. The effectiveness of proposed technique is presented with case and usability studies performed using generated and real-life datasets. In particular the Enron email dataset (323,073 emails, 19,898 email addresses over four years) and 20 Newsgroups (44,797 postings, 20 news groups and 5417 email addresses over one month) datasets are analyzed.

ZAMANA DAYALI BÜYÜK SOSYAL AĞ VERISETLERININ GÖRSELLEŞTIRILMESI

Uraz Cengiz TURKER

CS, Master Tezi, 2008

Tez Danışmanı: Yar. Doç. Dr. Selim Saffet Balcisoy

Anahtar kelimeler:Bilgi görüntüleme, Hiperbolik Yerleşke,Sosyal Ağlar, Anormallik Tespiti.

Özet

Sosyal ağ verileri sosyologların "Sosyal Yapı" olarak adlandırdıkları, insanlar arasında fikir ve bilgi aktarımını sağlayan, zaman ile evrimleşen haberleşme yapılarından oluşur. Sosyal ağ analizi ise özel hesaplama yolları ile bu yapıların özelliklerini ortaya çıkartır. Mevcut verisetleri milyonlarca insan ve bu insanlar arasında oluşan milyonlarca iletişim ağlarına sahip olabilmelerine karşın mevcut bilgi görüntüleme sistemleri sadece birkaç bin sosyal aktörülük zamana dayalı verisetlerini görselleştirebilmektedirler.

Bu tez, zamana dayalı büyük sosyal ağ verisetleri ile ilgili kavramsal çalışmayı ve zamana dayalı sosyal ağ veri setlerini interaktif bir şekilde gösterebilen yeni bir görselleştime metodunu anlatmaktadır. Önerilen teknik ideal gas yasası (IGY) ile Jacob Moreno'nun the Canon of Creativity teorisini 3-boyutlu hiperbolik ortamda sosyal ağ verisetlerini göstermek için birleştirmiştir ve 50,000 adet sosyal aktörlü bir veri setini interaktif hızlarda gösterebilmiştir. Konseptin doğruluğunu ispat etmek amacı ile sunulan methot kullanılarak, kullanıcıların sosyal ağ verisetleri ile etkileşimlerini sağlayan bir program geliştirilmiştir. Kullanıcılar mevcut ağı araştırabilir, görsel kalabalıklığı azaltabilir ve iletişimlerde oluşan anormallikleri gerçek zamanlı görebilirler. Bunula birlikte, kullanıcılar belli bir sosyal aktörü isleyebilir ve iletişim kanallarını görebilirler. Metodumuzun kullanılabilirliğini farklı şartları oluşturarak ve kullanıcı testleri ile sentetik ve gerçek veri setleri kullanarak gösterdik. Özellikle Enron elektronik posta veri seti (4 senede oluşan 323,073 e-posta, 19,898 e-posta adresi) ve 20 Newsgroups veri seti (1 ayda oluşan 44,797 mesaj, 20 news groubu ve 5417 e-posta adresi) analiz edildi.

1.2 Contribution

Temporal visualization of large datasets introduces new problems like "how to represent changes in the dataset by protecting clarity and user perception?" or "What is the interaction method?" Our proposed solution, which collects theories from sociology, mathematics, physics and geometry, is based on Moreno's theory of the Cannon of Creativity which we use as a **model** for revealing important social actors. Formulations of the Ideal Gas Law are used to construct the model, and hyperbolic space is used to support focus-context viewing.

The contribution of this work is not a graph drawing method but a novel representation technique for large temporal and non-temporal social networks datasets in 3D hyperbolic space based on the following important InfoVis criteria: Clarity, Scalability .

Researchers can incorporate our technique into visualization toolkits to analyze large temporal social network datasets easily. The main contributions of this thesis can be summarized as follows:

1) Pressure Model for temporal datasets: With the idea of the ideal gas law and the Cannon of Creativity, large temporal social network datasets are visualized.

2) Pressure Model for non-temporal datasets: Our method is also capable of laying out large non-temporal social network datasets according to SNA metrics in a 3D hyperbolic space.

3) Automated Visual Clutter Reduction: Our pressure model introduces a new visual clutter reduction method that is easy to understand and implement.

1. Introduction

1.1 Motivation

This thesis focuses on computer based visualization of large temporal social network datasets, which is a challenging research topic in itself. As social structures are composed through the interaction and communication between, it is important to consider time references to understand a social structure. Accumulating historical social network analysis data provides users to see the all of the communicative structure that occurred during a particular time frame.

For instance a scenario might be such as the following: During time *t* node *k* forms a relationship with node *i* once and node *j* forms a relation with node *i* many times and at time *t*+1 only node *i* forms a connection with *k* . If all historical data were to be erased, at time *t*+1 one could think that node *i* and *k* are closely related to each other, and thus should one position them solely according to the relational properties between them without concerning historical data, false topologies might be composed. As a result, to understand the meaning of all relational interactions, the storing and the processing of historical data is helpful.

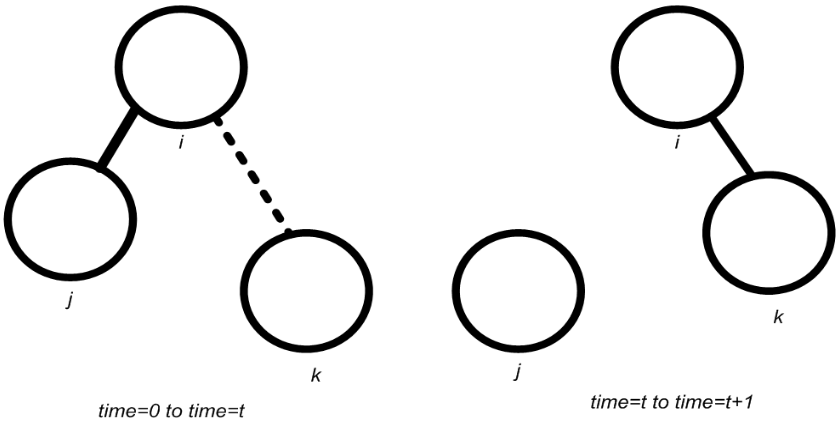


Fig 1.1 Dense relation between i and j can be observed only by accumulating past data between 0 to *t*

Although today's datasets can reach up to millions of items with many millions of relations with large temporal references, there are some physical limitations that prevent researchers to develop sophisticated visualization systems. For example, a screen size of a moderate monitor (res1600*1200) can represent at most 2 millions of items at a time. Besides, non-physical limitations also exist, not the least of which is the weak perception system of humans. The human visual perception system is able to perceive only 6 or 7 items among several items by comparing sizes, colours, and motions at a glance. Consequently, any visual representation of a dataset should exploit available system sources and must protect human perception in order to reduce the work load of the human cognitive system. As discussed in [6][8] visual perception depends on the clarity applied by the given layout. Clarity can be established by providing aesthetic rules like evenly distributed vertices, uniform edge lengths, and minimal edge crossings. Moreover, as depicted by Cruz [7] constraints like "placing important nodes near the canter of the display space" or "keeping a group of nodes close to each other" can affect the clarity in a positive way.

Computer oriented information visualization is a base between implicit information that is extracted from datasets and the human visual perceptive-cognitive system. InfoVis gathers theories and hypotheses from different fields of cultural studies like psychology, semiotics, and cartography. Physiologists study the human cognitive system, which transfers visual image into information. The research area of semiology is to find relations between meaning and representation. Cartography is a field of cultural study which abstracts meanings in terms of visual representations.

The problem this thesis focuses on is to represent valuable information, which is extracted from any temporal dataset, to domain experts, who have little or no knowledge about computer systems and computer programming, through the usage images.

We evaluated the developed system quantitatively with synthetic and real-life datasets. The tests with generated data prove that our methodology can represent social networks of various characteristics. We conducted an informal usability study with real-life datasets and found that users can perform quantitative analysis tasks with accuracy in an efficient way.

1.3 THESIS ORGANIZATION

In Chapter 2 related works on visualization of temporal and non-temporal social networks is discussed. Chapter 3 discusses the Sociology and the theory of the Cannon of The Creativity together with the Social Network analysis metrics. The formulation and the theory of combined gas laws are discussed in chapter 4. Mathematical explanation of different manifolds and detailed discussion about the hyperbolic geometry with isometries are presented in the 5th chapter. In chapter 6 we dive into the details of the proposed visualization system. In chapter 7 we show how information density can be increased by applying anomaly detection algorithms that we developed. Case studies and demonstrations of real-life datasets are presented and discussed in chapter 8. Explanation of the implementation and usability studies are in chapter 9. And finally we conclude by discussing the system and explaining the future works in chapter 10.

2 Related Work

Except some Geographical and Cluster Visualization systems, many of the temporal visualization methods, which support large datasets, use existing force directed methods to position nodes in displaced space. In addition, they mostly render edges between nodes that decrease the available display space and increase the visual clutter [8]. Furthermore, in their animations they rearrange the whole topology of the layout when the time slice alters and new is data loaded [40]. Rearranging the whole layout may disturb a user's perception, since a node may appear in any location of the display.

2.1 Visualizing non-temporal social network datasets

Analysis of social networks is an active research area in sociology. Moreno proposed one of the first visualizations for social networks, where actors are represented as circles and relations were represented with lines [20].
Mathematicians developed topologies to represent social networks using graphs in the 1970s [2]. Kamada and S. Kawai's force directed method, is one of the most powerful algorithms that visualizes social networks by solving force equations until the system achieves a stable condition. However, because of the force transfer calculations, early versions of force directed methods are only capable of visualizing small datasets. Later, enhanced methods are proposed that are capable of visualizing larger graphs, the maximum size not being more than ~20,000 nodes.

2.1.1 Physical Analogies

There are a number of social network analysis techniques that deal with static datasets. Thus, the software Pajek [41] is used in visualizing large datasets since it provides a stable and effective user interface. The Cavalier [38] toolkit incorporates several layout techniques; two important ones being spring embedder's [2, 3] and Kohen neural networks. We suggest the following references for a detailed research on static social network data exploration tools [3, 6].

2.1.2 Focus Context Techniques

Representing large information in a limited screen space is an important problem of information visualization systems. Focus context techniques are developed to represent large static datasets by manipulating the quantitative properties of visual variables like size and position, through either preserving the topology of the layout or not. The fish-eye [37] and the heat model [36] are two focus-context tools that alter the size of visual variables to establish focus-context images through preserving the topology of the layout.

Phillips, from the geometry centre [10] explained how to model a Hyperbolic Space by the usage of the Klein Model with efficient matrix operations. Tamara Munzner [11] showed one of the best graph drawing algorithms by which large graphs can be visualized. Although, the method is quite effective, it has some shortcomings as well. First H3 Viewer requires memory to represent large datasets. Second, User can get easily get disoriented while exploring the dataset.

2.2 Visualizing temporal social network datasets

Embedding an abstract dataset in a display space has been discussed throughout decades by researchers and as discussed above various techniques have been implemented to address the problem. However, very little research has been done to visualize the evolution of social networks [15,16]. As discussed by Müller [18] temporal datasets can be represented through static or dynamic techniques. Static techniques aim to show the condition and relation between members of the network with static images for a given time *t*. Dynamic techniques visualize the conditions of the network via animations. For dynamic techniques, however, since the topology of the layout changes with time visual complexity tends to occur and this is an important challenge to overcome [6]. SoNIA [60] is a graph animation tool that generates videos of animated graphs based on different layouts with little interactivity. PieSpy can visualize dynamic data [43]. Kumar and Garland [15] presented a hierarchical force-directed layout technique, which addresses the perception issues successfully. However, this technique requires a priori knowledge of graph hierarchy. Condor [16] has a new algorithm called *Sliding Frame* to represent the evolution of the dynamics of a social network through time. Again these methods utilize the force directed method in laying out the dataset leading to visual-computational complexity issues [15,18].

2.3 Comparison with State Of the Art

Most social network analysis methods aim to represent dataset as graphs that represent relations. However, the sole aim of these representations is to convey the structure of the social network. Our framework generates clear and self expressive representations of real life datasets based on social network analysis metrics, and the advantage of our method over existing tools are as the following:

1) Understand and see the evolution of a very large **temporal** social structure of **more than** 50.000 nodes at interactive speed.
2) Real time interaction with a **temporal** dataset to focus on a social actor or group and follow his/their social communication channels.
3) Analyze and investigate any personal/group/structural abnormal relations.
4) Unlike others, the proposed system can render 50.000 nodes with 350,000 edges on systems that have only 128 MB of memory.

3 Sociology

Sociology is a field of social science which studies the behaviour and intercommunicative structure of a society, the effects that manipulate the structure of a society, the evolution of societies, properties of personal relations, and other topics that concern the structure of intercommunication among people. The foundations of sociology enable people to understand and govern societies [20]. In [21] Jacob Moreno discussed the sociological background of the German nation and he proposed sending sociologist to Germany in order to eliminate physiological conditions that raise hostility against the USA after WW2. Furthermore sociologists studied the past behaviours of Soviet society and Soviet governments to predict possible diplomatic steps in the Cuban missile crisis during the Cold War to aid the Ministry of Foreign Affairs of the USA [25]

3.1 The Cannon of Creativity

Jacob Moreno is the founder of modern sociometry and psychodrama. In his working he discussed the philosophy behind human social activities and aimed to cure patients through social activities. This notion has indeed led to the concept of Group Therapy in later years. In his book entitled *Who Shall Survive?* Moreno explains his elaborate studies concerning social activities and social structures between individuals and points out that there is a similarity in between biological space and social space in terms of Darwin's theory of The Natural Selection.

Based on Darwin's theory weak biological actors cannot support their basic needs and cannot survive and cannot have any descendants. They will thus become extinct and there will be no evidence left of their existence. Moreno states that in a social cosmos, weak social actors are known by a small group of people such as their families, friends and colleagues. Conversely, important social actors are known by masses. For example; although there were numerous people with names such as "Leonardo", "Newton", and "Galileo" in the past; today one only remembers "Leonardo de Vinci", "Isaac Newton", and "Galileo Galilee". He focused on this phenomenon and inferred that social activities determine the duration of presence of a social actor into the future. Finally he ended up with the theory of the *Canon of Creativity*. In short Moreno states that a social actor must perform social activities to adapt to or to survive in a society. The rank of an actor is improved by his/her social activities. The higher the number of social activities, a social unit preserves or enhances its position and becomes more important.

The Cannon of Creativity is a formulation for a social cosmos, which reveals the importance of social actors based on four distinct intrinsic properties; "spontaneity", "creativity, "action", and "preserve". These properties do not have any interconnection between them. Thus being spontaneous does not mean that a person is also creative, being creative does not mean that a person can act and perform her ideas, and lastly being active does not mean that all actions are creative so that she preserves. A strong social actor has to have enough time to act and has to be creative. Beethoven had some special abilities; he was creative, and he had enough time to write great symphonies. Michelangelo had as much time to paint as many other artists of his day, but his creativity separated him from others and made him important social actor.

To express the theory of the Cannon of Creativity I will present two examples; Hazerfan Ahmet Çelebi (Çelebi) and the Wright brothers. These people were able to create machines which could fly, however the answers of why they haven't equal social importance, is the difference between their spontaneity, creativity and act.

Spontaneity: Is the temporal ability (energy) of a person which yields social activity. In short; it is the free time. If one has enough free time, one can think about new ideas and one can try to realize these ideas. The old analogy is the light. If light is turned on, the objects in a room become visible and usable but if light is turned off, although the same objects remain in the same room, they become invisible and useless [20].

Çelebi and the Wright brothers had enough time to think to create devices that enable humans to fly. However the duration of their spontaneities was different. After his flight, Çelebi was exiled to Algeria by the Sultan as a punishment; but the Wright Brothers had enough time to get a patent and founded the Wright Company before they died.

Creativity: in terms of sociology it means fertile thinking. By ‘fertile’ we mean an output of an idea which has a purpose. Creativity is measured according to the usage of its outcome.

The creativity of Çelebi is weaker then that of the Wright brothers’ because, Çelebi is the first person who thought to fly across the Bosporus in Istanbul by utilizing the physical force of a human being only. The Wright Brothers, on the other hand, were the ones that thought to create a plane which has an engine that is capable of carrying people and goods.

Act: Realizing the idea. Action is the output of creativity but again it requires spontaneity.

Çelebi built two wings and jumped from the top of the Galata tower in Istanbul, flew across the Bosporus and landed on Dogancilar Square, Üsküdar (6000 m). The Wright brothers built their first plane with stationary wings and carried one passenger and lifted it up to 107 meters.

Preserve: is how strong the social character is, the number of people who know a specific character. Preserve is the strength of the social importance.

If we compare the number of people who know Çelebi with the number of people who know the Wright brothers, it is quite evident that the Wright Brothers are known by a much larger group of people. Because, Çelebi thought to fly across the Bosporus; this idea (creativity) is local in two senses. First Çelebi’s aim was to fly, second he thought to fly across the Bosporus only. Moreover, he had no opportunity for spontaneity after the Sultan’s punishment. However, the aim of Wright Brothers was not local, their ideas and purposes targeted to establish a norm to carry people from one point to other. Their ideas affected a large number of people. Furthermore they had enough subsequent spontaneity which provided them with an opportunity to enhance their works.

3.2 Social Network Analysis

Social network analysis is a field of sociology that aims to extract structural information of a social structure from a social network dataset, and it explains the behaviour of a social structure through mathematics and graph-theory [19]. Social network analysis provides visual and mathematical analysis of social space.

Because, the behaviour and actions of a social community are characterized and applied by their important social actors, in social network analysis the identification of these important actors has a top priority [19]. For example; Valdivs Kreb collected information from publicly available sources (Online Newspapers, Reports, etc.) and applied social network analysis which revealed that Mohammed Atta is the central person in the terrorist group, that is responsible for the 9/11 attack. The same analysis also reveals the hidden relation between Al-Qaida and Mohammed Atta. Likewise, Google’s page-rank system applies social network analysis to order pages according to a popularity metric to detect trends and apply intelligent-advertisement.

Social network analysis applies graph-theoretic functions and researchers use graphs to represent the qualitative properties of social structures. A mathematical definition of a graph is the following, “a graph consists of finite set of vertices some pairs of which are adjacent, joined by an edge. No edge joint a vertex by itself and at most one edge joins any two vertices”. [26] The mental mapping of this definition with any domain is straightforward: A node can be an abstraction for any entity and an edge can be used for any kind of relation. In terms of computer networks, nodes represent e-mail addresses and edges represent any communication between e-mail addresses. In the investigation citation datasets, nodes may imply academic papers and edges may represent issued references. Similarly, from a sociological point of view; node-link diagrams are good abstractions for social structures where nodes represent social actors and any relation between individuals is represented as an edge. Below there is the node edge representation of the relations between the engineers in an office. The discussion of social network measures will be done through this example.

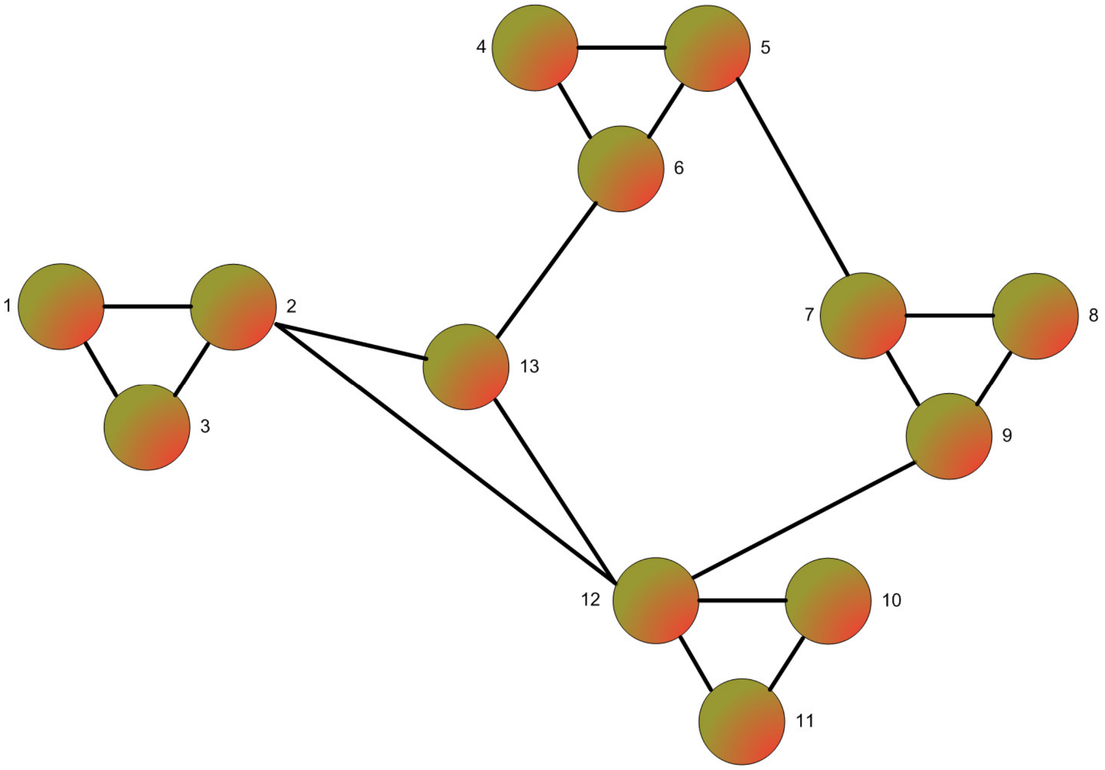


Fig 3.1. Sample Social Structure

3.2.1 Degree Centrality

Degree centrality is the measure of the direct connections of a social actor with other social actors in a given social structure. Although its computation is the simplest among several SNA metrics, it is still very important. For instance; in terms of medical discourse; degree centrality can be applied to a person who is infected by a virus to figure out his/her direct relations to control the expansion of the virus. There are some circumstances where these relations have directions such as citations. For those directional centralities; *in-degree centrality* and *out-degree centrality* can be calculated. Calculation of the degree of centrality is as follows:

$$N_i = \frac{\sum 1}{g-1} \qquad (1)$$

Where *l* is the total number of connections of node *i* with other nodes; and g is the total number of nodes in the graph. In the above figure, the calculation of degree centrality reveals that node12 is a central person and has more direct connections.

3.2.2 Betweenness Centrality

Betweenness centrality is the reachability of a social actor by other social actors. A social actor's betweenness centrality is high if and only if, the number of social actors, who can reach other social actors through shortest paths passing through the social actor, is high. Today, the USA's armed forces widely use betweenness centrality on convicted persons to reveal other suspects quickly.

$$N_i = \sum_{s \neq v \neq t \in V} \frac{\sigma_{\mathrm{st}}(v)}{\sigma_{\mathrm{st}}} \qquad (2)$$

v is a vertex, notation σrefers to the shortest path which is between the distinct vertices *s* and *t,* and *V* is the list of the vertices in the social structure. According to the betweenness centrality the number of shortest paths passing through node12 is the largest.

3.2.3 Closeness Centrality

Closeness centrality is used to obtain social actors who are close to everyone. The term *close* refers to the shortest path. So closeness centrality is the measure of the number of shortest paths of a social actor to all social actors in the social structure. A close social actor hasaccess to information with ease within the social structure. It can be calculated by taking the inverse of the sum of all shortest paths to others:

$$\mathrm{Ni} = \left\{ \sum_{j=0}^{j=g-1} \sigma_{ij} \right\}^{-1} \qquad (3)$$

The application of this formula to our sample office dataset results that node12 has the ability to reach every other actors very fast.

3.2.4 Network Centrality

Network centrality is the structural property of a social structure that infers the distribution of the social structure. A very high central structure is dominated by few but powerful (very high degree centrality) social actors, which hold other social actors together and if the central actors are down, the social structure fragments into small social groups or collapses. It can be seen that the above example has several central nodes, like node2, node13, and node12.

4 Ideal Gas Law

Ideal Gas Law is the mathematical formulation that describes (approximates) the behaviour of gas molecules that are hypothetically in an ideal state. In physics, the Ideal Gas Law (IGL) is used for several real life problems, such as analyzing and separating mixed gases which reside in the same environment, with respect to their distinguishing properties through following equation [17].

$$PV = nRT \qquad (4)$$

Where *P* is the pressure of the gas, *V* is the volume that gas occupies, n is the number of molecules; *R* is the universal gas constant, and *T* is the temperature of the environment. The formulation is the most accurate for the monatomic gases because ideal gas formulation neglects the intermolecular attractions. The formulation of ideal gas law is based up on several gas laws;

4.1 Charles' law

Jacques Charles proved that if the mass of a gas remains constant in a closed environment with constant pressure (isobaric process), necessary volume for the gas varies directly with the absolute temperature.

$$V{\sim}T \qquad (5)$$

According to kinetic energy theory and the conservation of energy, speed of a particle is referred as its temperature with the given relation,

$$\frac{1}{2}mS^2 = T \qquad (6)$$

Where *m* is the mass *S* is the velocity of the molecule, and if the equality is applied in ideal gas law formula than we obtain;

$$PV=nR[\frac{1}{2}m\, S^2] \qquad (7)$$

After eliminating P, n and R we obtain the lemma; the volume of a molecule is directly related with the kinetic energy of a molecule.

$$V \approx [\frac{1}{2}m\, S^2] \qquad (8)$$

4.2 Boyle's law

Boyle's law states that the pressure and the volume of a gas are inversely proportional if and only if the mass of the gas and space temperature are constant.

$$P \sim 1/V \qquad (9)$$

$$P_{t1}V_{t1} = P_{t2}V_{t2} \qquad (10)$$

P_{t1} and V_{t1} refer pressure and volume at time *t1* respectively, and P_{t2} and V_{t2} refer pressure and volume at time *t2* where *t2>t1* and $\{V_{t1} < V_{t2}$ OR $V_{t1} > V_{t2}\}$. The proof of this theory can be obtained by kinetic theory. Any volumetric alternation of the space affects the speed of molecules and the amount of elastic collusions inside the closed space so the pressure *P* alters.

4.3 Gay-Lussac's Law

In his working; Joseph Lois Gay Lussac [48] proved that under constant temperature; the pressure of a gas is inversely related with the volume but is directly related with the temperature.

$$P \sim T \qquad (11)$$

According to kinetic theory and the formula (7), as the kinetic energy of molecules increase, the amount of collusions inside a space increase, consequently pressure of the space rises. Similarly, if the volume of a gas decreases, statistically rate of collusions increases.

4.4 Avogadro's law

According to Avogadro's law; under constant pressure and temperature, equal volumes of gases contain same number of molecules with the following equation:

$$(P_{t1}V_{t1})/\,(T_{t1}) = (P_{t2}V_{t2})/\,(T_{t2}) = c \qquad (12)$$

Where c is a constant $P_{t1,}$ $n_{t1,}$ T_{t1},and V_{t1} refer pressure, number of molecules, temperature, and volume at time *t1* respectively, and $P_{t2,}$ $n_{t2,}$ T_{t2},and V_{t2} refer pressure, number of molecules, temperature, and volume at time where *t2>t1* and $\{V_{t1} < V_{t2}$ OR $V_{t1} > V_{t2}\}$, and $\{P_{t1} = P_{t2}$ AND $T_{t1} = T_{t2}\}$. The constant value c is referred as Avogadro's number = 6.022×10^{23} particles per mole if and only if V = 22.4L T = 273K and P = 100bar

5 Geometry

Geometry is a field of mathematics which focuses on the quantitative properties like lengths, sizes, areas, volumes, and positions of topologies in a given manifold where a manifold is a mathematical abstraction for an n-dimensional space, in which Euclidean postulates are applicable [49]. For example; imagine a line *L* lies on an axis of a very large spherical space as in figure.

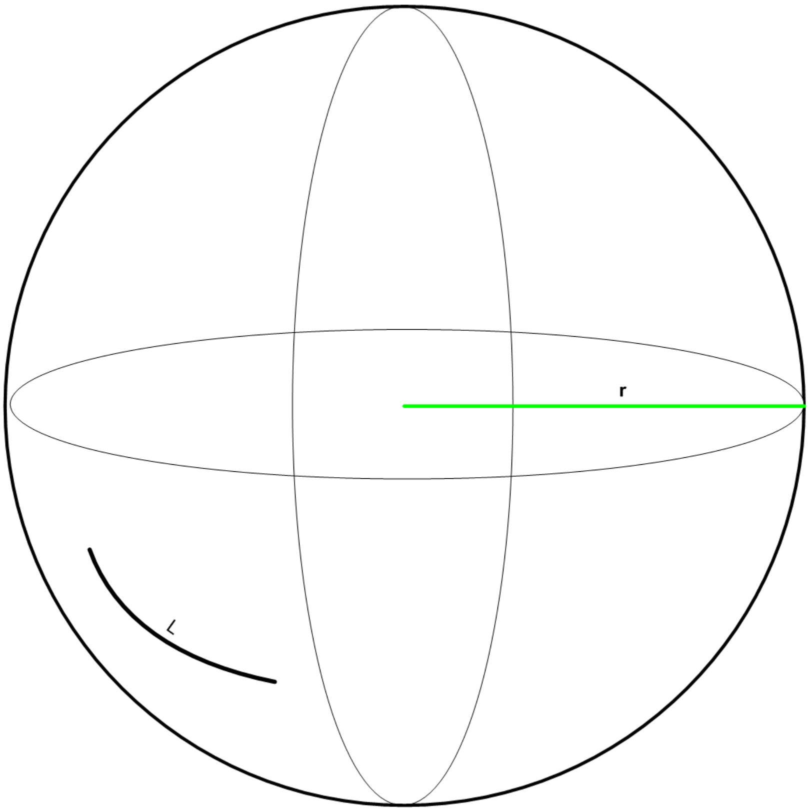


Fig 5.1. Line *L* lies on a Spherical Surface of radius r.

We can observe that if the length of this line gets smaller the line seems so straight that we can draw a solid triangle using this line and apply Euclidean Geometry. This is exactly the same when we perceive the world as flat when we look around in streets.

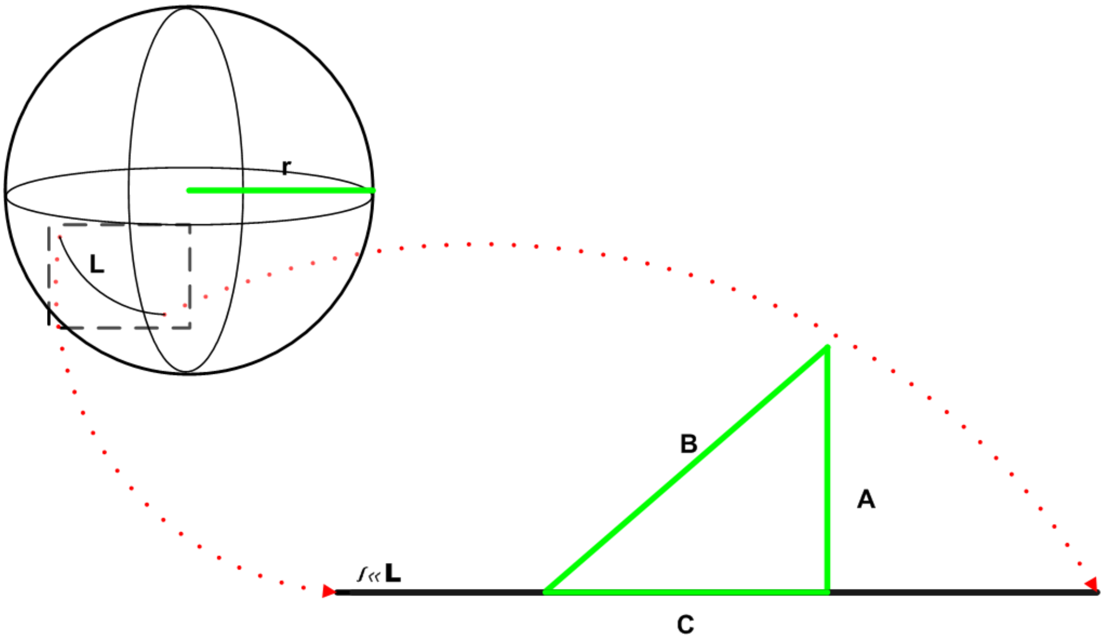


Fig 5.2 As the length of the line L approaches to *l = 0* the geometric properties of line L approximates the Euclidean postulates.

Moreover, according to Einstein's relativity theory our world and universe is curved that obeys some n-dimensional manifold geometry, but geometrical properties of Euclidean Space are still valid.

Classification of 3D manifold geometries is an active area of mathematics, but because the discussion of this thesis is not geometry and properties of manifolds, widely known geometries will be discussed shortly. Here are the definitions of some geometric primitives and terms:

Point: a unit element of a space in n-dimension.

Line: summations of infinite points, which lie on the shortest path, between two distinct points.

Lie on: distribution of points of an topology on a n-dimensional surface

Between: is the condition of having points that are member of the shortest path elements passes two points.

Congruent: if an object *b* can be obtained by isometric transformations of object *a* these two objects are congruent.

Plane: summations of all parallel lines lie on the shortest path between two points

Space: boundles n-dimensional continuum

Shortest path: Is the path between two points with the minimum number of points, or the minimum value of the distance value according to the manifold metric.

Straight Line: Composition of a path between two points where the cross product of every adjacency points is 0.

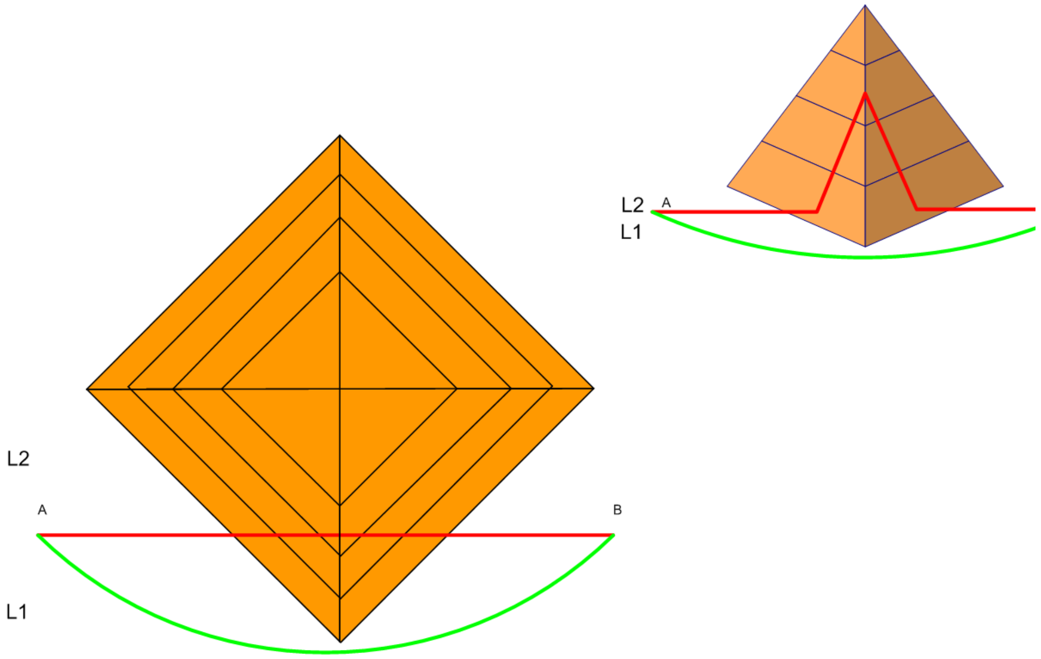


Fig 5.3 Note that it is not necessary to use straight lines as shortest paths, Geodesics are shortest paths of curved spaces

5.1 Euclidean Geometry

Euclidean geometry is widely known, and was mathematically proved by Euclid. In Euclidean geometry lines and planes are flat; so summations of any shortest path between two points said to be *straight*. Lengths, sizes and areas are formed around Euclid’s five postulates:

1) A line segment can be extended to infinity
2) If a shape can be obtained by another one through applying an isometry, these two items are called as congruent.

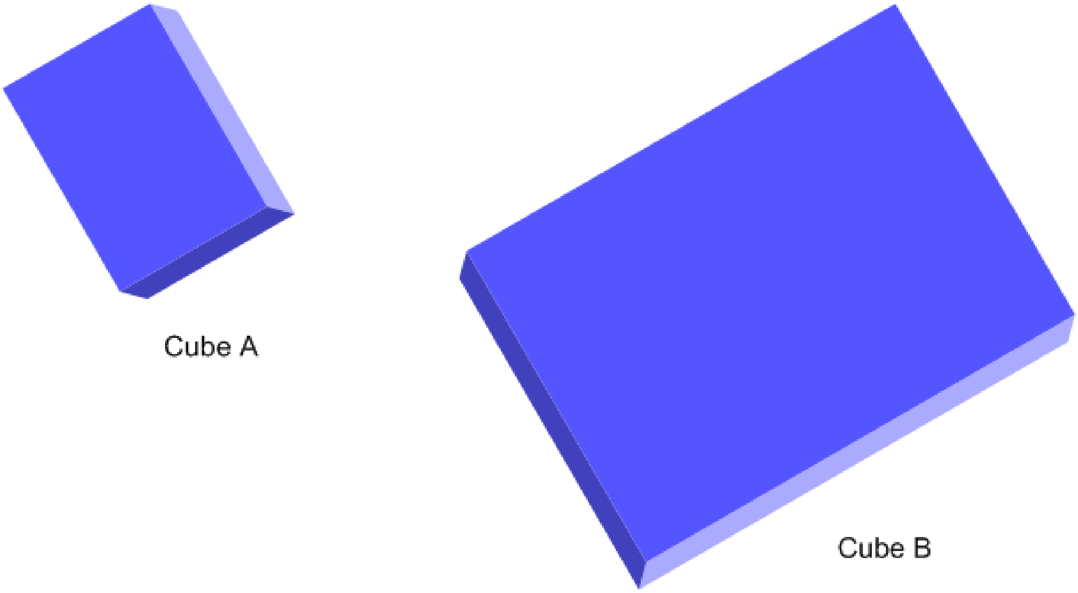


Fig 5.4 Two congruent boxes. Rotation around X axis and 3D scale result the image at the right.

3) A line can be drawn to connect two distinct points.
4) A line can be a radius of a circle

5) In Euclidean geometry if two lines are parallel the distance *d* between these two lines is never change and there is only one line is parallel to another line that passes through P

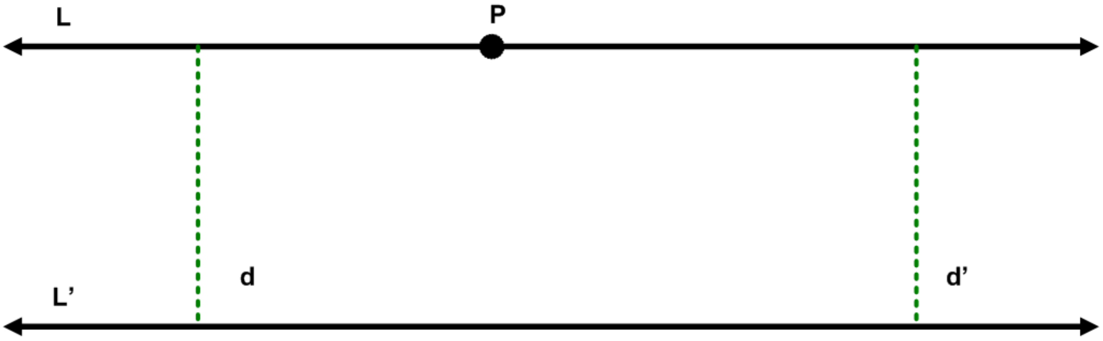


Fig 5.5 Two parallel lines in Euclidean Space where d= d'

Many information visualization techniques tend to layout large graphs in Euclidean space by applying abstractions, or focus-context techniques. Abstractions are done via encapsulating the nodes by isosurfaces. These isosurfaces have physical properties like colour, size, and motion to give information about their contents. However, for large datasets these abstractions prevent users to see the whole dataset and relations at the same time, hence information retrieval cannot be supported adequately. Focus-Context techniques are useful for large graphs to dive into the details of the graph. However, with these techniques users can lose their sense of coherence regarding the topology of the graph.

The main reason to use abstractions or focus-context techniques is the polynomiality. In Euclidean space the area of a simple uniform square where the side lengths is *a,* is calculated with the polynomial a^2 and the distance between points is given by:

$$d(\mathrm{a,b}) = \{(a_1-b_1)+ (a_2-b_2)+\ldots\ldots\ldots+ (a_n-b_n)\}^{1/2} \qquad (13)$$

However, when considering social relations the growth rate of siblings of a node are exponential. For instance; if a node *i* has *m* numbers of siblings at level *k*, it will have $m^{(k+3)}$ siblings at level *k*+3.

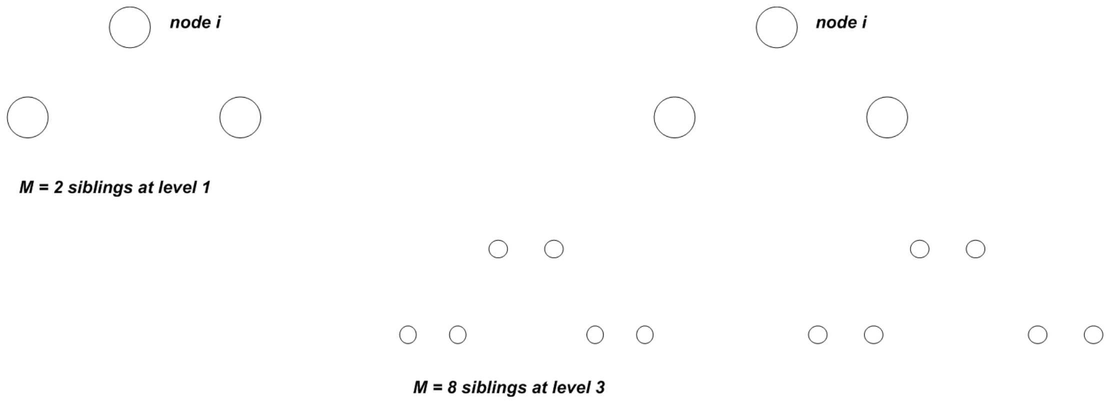


Fig 5.6 leafs of the tree increases exponentially

5.2 Spherical Geometry

Spherical geometry is a geometry, applied on a surface of a sphere where the sphere is an isosurface composed by points in three-dimensional space that are equidistant from a certain fixed point given by the polynomial;

$$f(x,y,z) = x^2 + y^2 + z^2 = 1 \tag{14}$$

Although Euclidean geometry is widely known and applied, because of the shape of the Earth, the first studied 3D manifold is a Spherical manifold geometry. Euclid’s parallel postulate is not applicable for Spherical Geometry, and actually there is no “parallelism” in spherical geometry. As the geodesic of a spherical space is called the great circle, there are no any parallel lines. The unique topology of a sphere results in the following rule: Given a line *l* and a point not on *l*, no lines exist that contains the point, and are parallel to *l* and all geodesics meet at two different points.

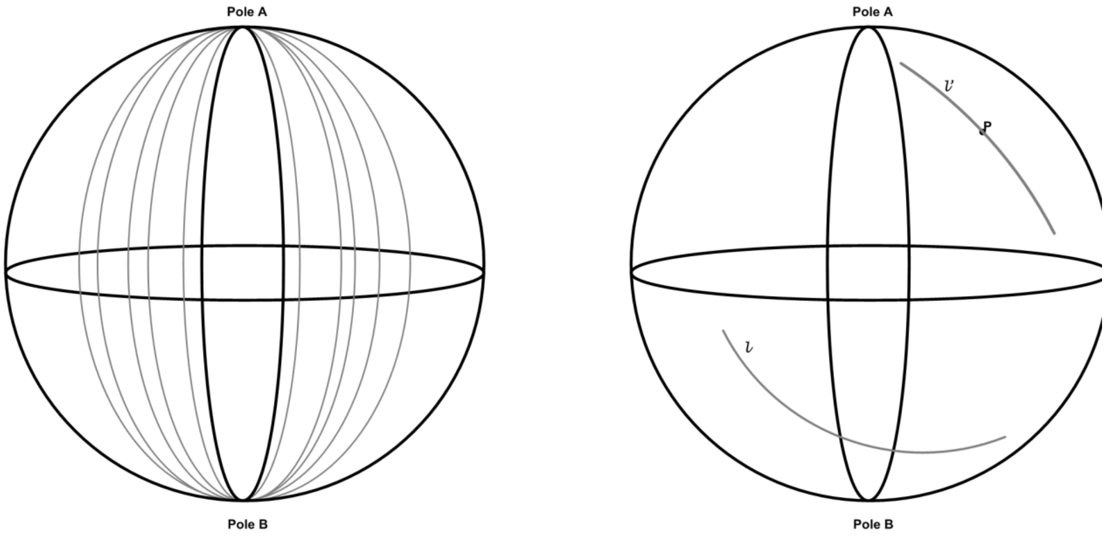


Fig 5.7 infinitely many different numbers meet at poles (left) there is no any parallelism (right)

Moreover as can be seen from Figure 5.7 Euclid's third postulate is violated since the poles can combine infinitely many numbers of distinct lines. In addition, the first postulate of Euclid is also violated because of the previous property.

Here are the postulates of the Spherical Geometry:

1) Two points lie on a unique straight line if and only if they are not antipodal. Infinitely many lines can pass through the Antipodes.

2) An unbounded line meets with itself when it has the length of the great circle.

3) The radius of a circle must be smaller then the half of the circumference.

4) Spherical Geometry is uniform and it holds the statement of uniformity. All right angles are equal.

5) All lines meet in two points.

Information visualization techniques layout datasets that utilize spherical geometry do exist [33]. In Radial layouts nodes are positioned on concentric circles according to their depth in the spanning tree and if a sub-tree exists, it is laid out over another circle. However spherical geometry does not address' the problem of polynomiality where the distance between two points *a b* is given by

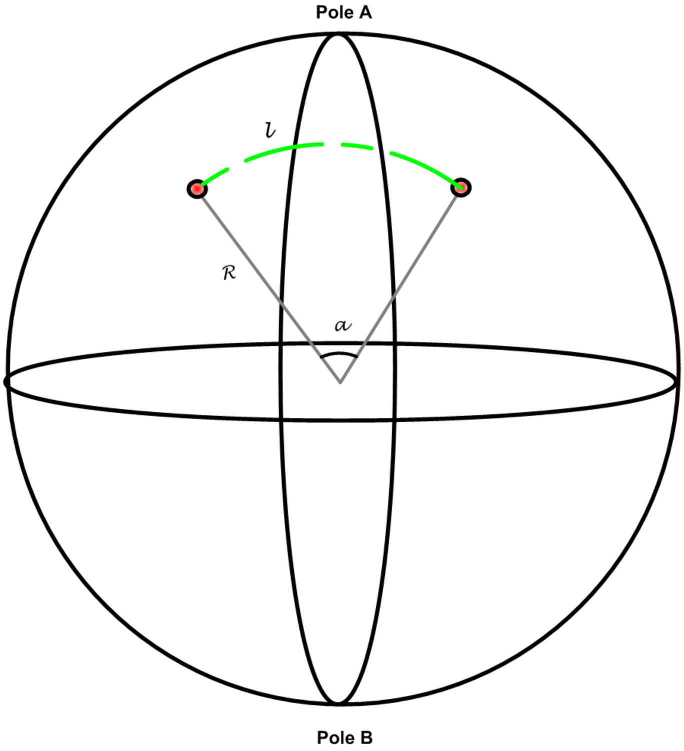


Fig. 5.8 distance between two points in spherical geometry

$$d(\mathrm{a,b}) = R,\hat{a} \qquad (15)$$

Where R is the radius of the great circle and $\hat{a}$ is the angle between two points.

5.3 Hyperbolic Geometry

According to Greenberg the answer to “What is hyperbolic geometry?” is “the honest answer is that we don’t know; it is just an abstraction” [58]. Hyperbolic geometry is one of the most useful and important kinds of non-Euclidean geometry. The discussion of hyperbolic space is closely related with the Minkowski Space. Minkowski space is a 4-D the space-time space as in the figure:

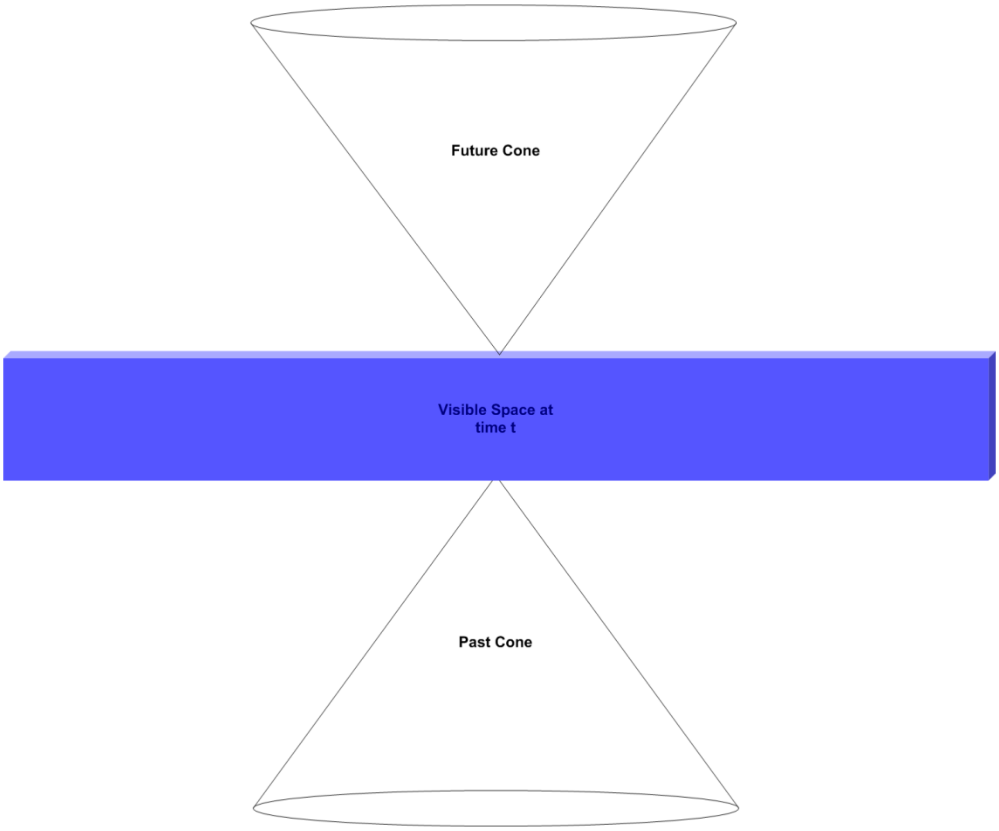


Fig 5.9 Minkowski space

According to the theorems of the Minkowski space and relativity, an action can not be observed by the whole space at the same time, rather it can be viewed in the future and it can be affected by the past actions in the same space because of the limitation of the speed of the light. This is an explanation of why we can see supernovas that have already happened many centuries ago, today. As in figure 5.9 this phenomenon divides the space-time space into two distinct conic sections. The upper side is an abstract space consisting of infinite observers who will observe the action performed at time *t*, where the lower side is an abstract space consisting of infinite actions that happened in the past and that can be observed at time *t*. The bounds of these conic sections are limited with the speed of light and proved by the Lorentz formulations.

In Minkowskian space, considering a set of points that have equal distances from a fixed point say point (0,0,0) forms two hyperboloids

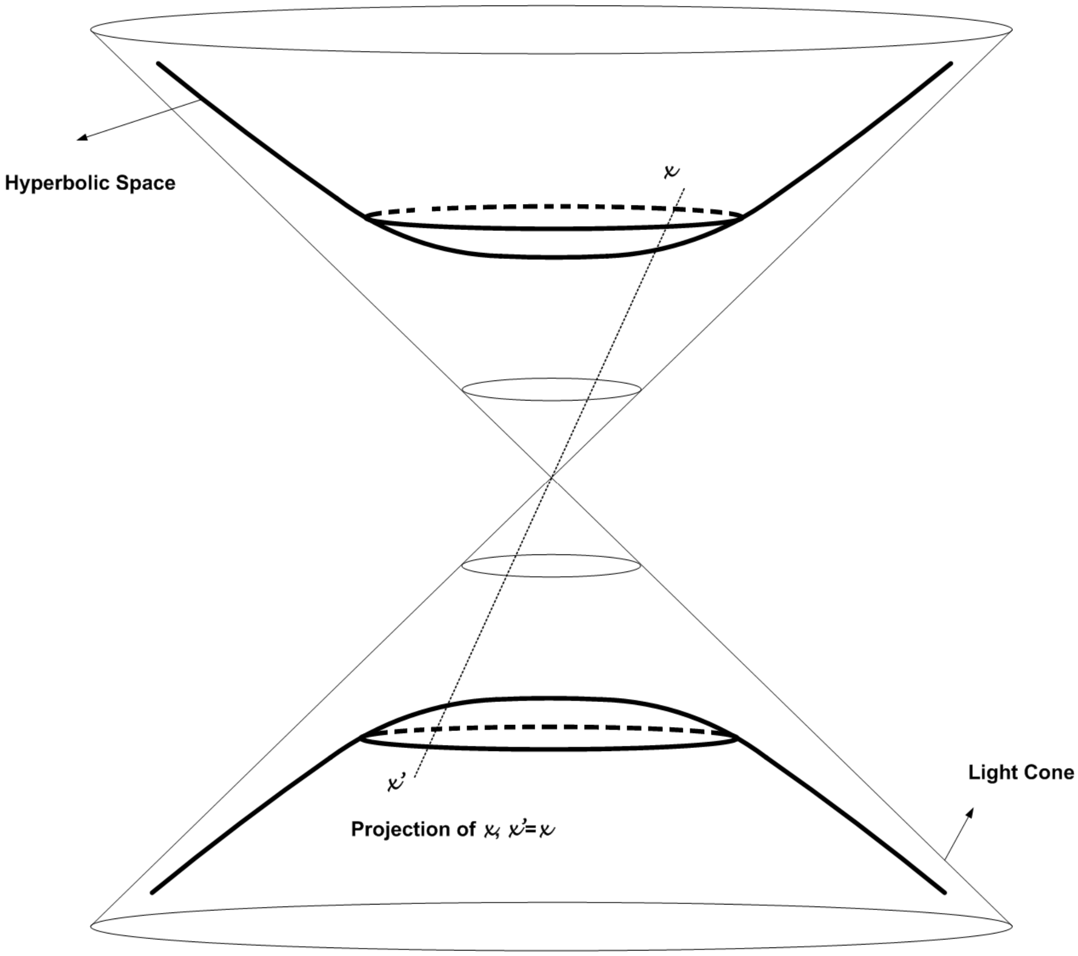


Fig 5.10 Hyperboloids

The derivation of hyperbolic space from Minkowski Space is similar to the derivation of a unit sphere in Euclidean space where the sum of all squared distances from a point is equal to 1

$$S \ \ \langle p.p \rangle = 1 \qquad \text{where } p = (\mathrm{x,y,z}) \text{ (in case of 3D)} \qquad (16)$$

Similarly, if we equate the sum of all squared distances to -1 we obtain a hyperbolic space;

$$\mathrm{H} \ \langle p * p \rangle = -1 \qquad (17)$$

Where * is an Minkowski inner product given by:

$$\langle x * y \rangle = \left(p_{x_1} * p_{x_2}\right) + \left(p_{y_1} * p_{y_2}\right) - \left(p_{z_1} * p_{z_2}\right) \qquad (18)$$

As a result of this property hyperbolic space can be mapped into a unit sphere in E^3 where the curvature is -1. Because in hyperbolic space Euclidean straight lines are curved as in figure 5.10, all Euclidean postulates, except the parallel postulate, are applicable to hyperbolic space.

The parallel postulate of Euclidean geometry states that the distance between two parallel lines is constant and there is only one line that lies on a point *P* that is also parallel to a line. However, in hyperbolic geometry that given line *L* and a point *P* not on *L*, there are an infinite number of lines passing through *P* parallel to *L*

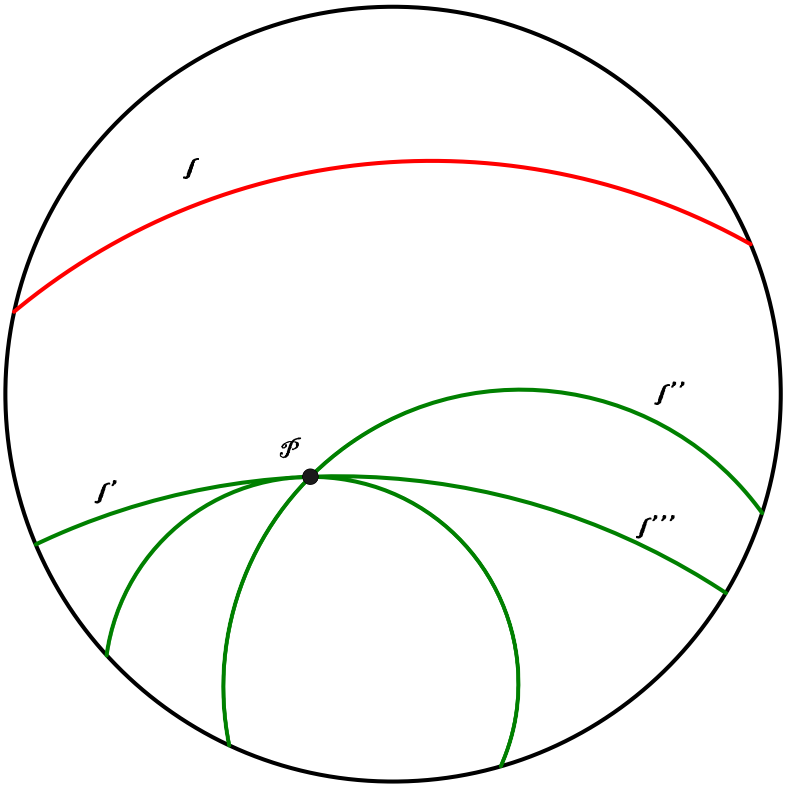


Fig 5.11 there are lines *l’, l’’, l’’’* are parallel to line *l* and passes from point *P*

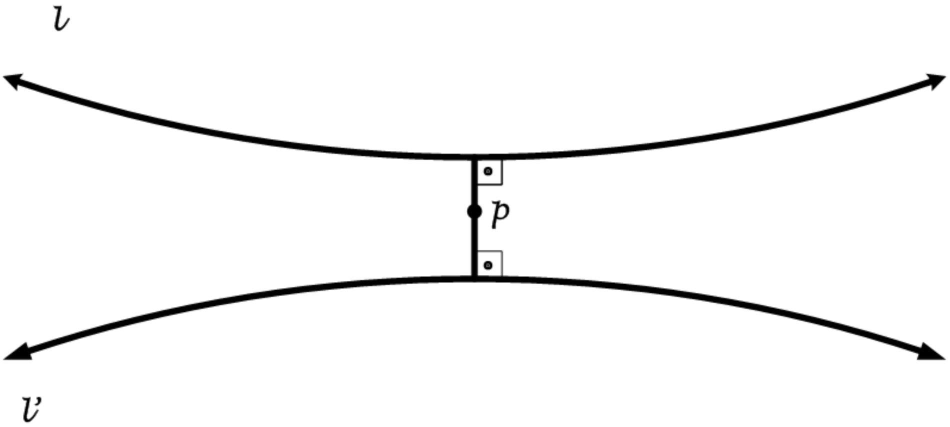


Fig5.12 Two parallel lines in Hyperbolic space, if the distance between two lines approaches to 0, lines do not unite.

Another important result of bent space is called *the defect* of hyperbolic space, where the cumulative values of interior angles of a triangle must be **less** than 180.

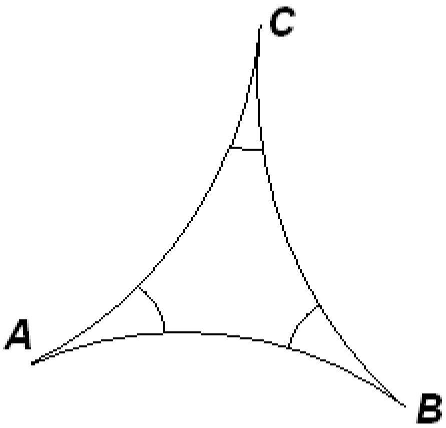


Fig5.13 A triangle with same angles CAB = ABC = BCA = 28° and the total cumulative angle is 84°

There are five analytical models whereby we can generate hyperbolic space which have their own isometries, metrics and calculations.

1) Hyperboloid Model
2) Klein Model
3) Poincare's Half Plane Model
4) Poincare Disk Model

At the beginning of our discussion on hyperbolic geometry we introduce the hyperboloid model however; because the proofs and explanations of all these models is not the scope of this thesis, only the Klein Model is elaborated in detail.

Klein Model

The Klein model is a projection of a one-side of the hyperboloid in to an *n-dimensional* Euclidean unit circle as in figure 5.15. Let *l* and *m* two distinct lines in upper side of the hyperboloid. And let *l'* and *m'* are the projections of these lines onto the 2D plane. Klein model is obtained by the transformations of these projections of hyperbolic lines onto the 2D circular plane. As a result the generation of the Klein model is done by the following step: F is a circle with origin O and radius *r*.

$$\mathrm{P} = \{X \text{ € } \overline{OX} < Or\} \tag{19}$$

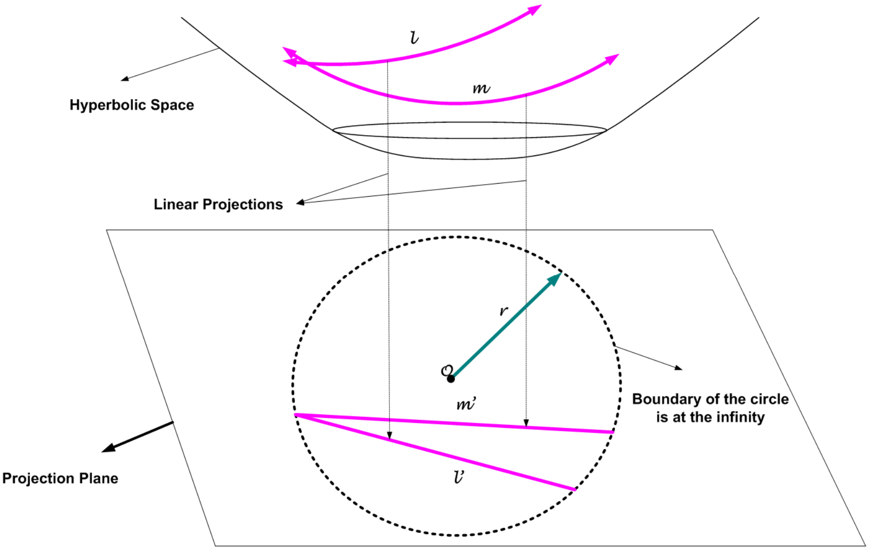


Fig 5.14 Projection from upper side hyperboloid to 2D plane

And with this class of points we say that a line of *l* that passes through points forming *l'* is an open line in the hyperbolic space. Although in the Klein model lines are Euclidean lines, angles between lines are not Euclidean angles.

5.4 Isometries

In this section three isometries in hyperbolic geometry; reflection, translation, and rotation will be explained as given in [10]. The Klein Model is also referred as the Projective Model projects entire hyperbolic space onto a unit sphere or circular plane where isometries can be represented as linear transformations. To implement these transformations on computer graphics hardware, which are optimized for 4 by 4 homogenous matrix operations, the conversion from affine coordinates to homogenous coordinates is vital and done by Plücker transform.

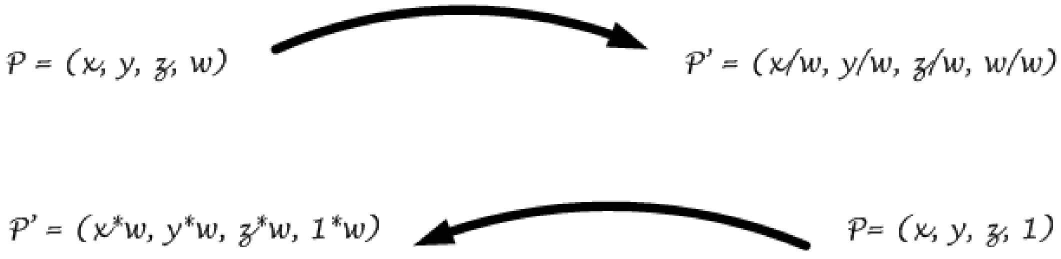


Fig 5.15 In computer graphics Plücker transformation used to convert affine coordinates to homogenous coordinates

Reflection is the process of reflecting all points according to a reference point P. In Klein model reflection can be given with the matrix formula;

$$R = I\text{-}2pp^{t}I^{3,1} /\langle p * p\rangle \qquad (20)$$

I is the identity matrix and $I^{3,1}$ is the Minkowski identity matrix:

$$I^{3,1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix} \qquad (21)$$

Translations of points from point *a* to point *b* can be given in terms of reflections

$$T_{a,b} = R_m.R_a \qquad (22)$$

R_m is the middle point between point *a* and point *b* given by

$$R_M = a\sqrt{\langle b.b\rangle\langle a.b\rangle} + b\sqrt{\langle a.a\rangle\langle a.b\rangle} \quad (23)$$

And "." is the Euclidean scalar product.

Rotation is the process of positioning all points according to a **reference point *P*** with a supplied angle θ. The **reference point *P*** must be lie on the rotation axis and is given by;

$$P = b[\frac{a.(a-b)}{((a-b).(a-b))}] + a[\frac{b.(b-a)}{((b-a).(b-a))}] \quad (24)$$

where *a* and *b* are affine coordinates and . is an Euclidean scalar product. In Klein model rotation can be given with the matrix formula;

$$R = {T_{lo,0}}^{-1} + R_{(u,\theta)} + T_{(lo,0)} \qquad (25)$$

$R_{(u,\theta)}$ is the rotation matrix that can be given by :

$$R(u,\theta)\begin{bmatrix} (x^2 + (C*(1-(x2))) & ((x*y*(Z)) - (z*S)) & ((x*z*(Z)) + (y*S)) & \mathbf{0} \\ (x*y*(Z) + (z*S)) & (y^2 + (C*(1-(y2))) & ((z*y*(Z)) - (x*S)) & \mathbf{0} \\ (x*z*(Z) - (y*S)) & (z*y*(Z)) + x*S) & (z^2 + (C*(1-(z^2))) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1} \end{bmatrix}$$

(26)

Where C = $cos(\theta)$ S = $sin(\theta)$ Z = $1 - cos(\theta)$. And x,y,z are the vector entries of the unit vector U;

$$U = (\frac{a-b}{\|a-b\|}) \qquad (27)$$

6 Pressure based layout method

Proposed method is based on ideal gas law, which is employed for social network analysis in line with the cannon of creativity theory in section 3. Theory of the canon of creativity can be summarized as: Supplying necessary communication provides continuity in terms of size and position in social space, which can be reformulated by the formulations of the Ideal Gas Law (IGL). From now on, we refer social actors as nodes, and we use terms node and gas interchangeably. In our IGL analogy the degree centrality measure of a node is said to be the number of molecules of a gas (n), and the size of the node is related with the volume of a gas (V), and summation of all degree centralities in the social network is the space pressure (P).

Similar to the Avogadro's law, where numbers of molecules are distinctive properties of a gas, in social space social activities are distinctive properties of social actors [20]. So, based on this similarity, unique position and representation of one social actor in a group can be calculated by IGL with the following rule of inference: *distinctive properties of a social actor can be obtained by Ideal Gas Law*.

Based on this, we derive four hypotheses:

1. The interior pressure of a node alters directly proportional to its centrality measure.
2. Representation of a node has to alter according to the ratio between interior and outer pressures.
3. Space pressure increases as the total value of the degree centralities of nodes inside the space increases.
4. Nodes with larger sizes tend to come to the center of the space. Weak nodes are pushed to distances.

The first three satisfy the following statement of Moreno: Supplying necessary social intercommunication provides continuity in terms of size and position in space. Forth hypothesis is an embodiment of the law of *social gravity* [20]. The law of social gravity is the ratio between attraction and repulsive forces among two social actors inversely proportional to the distance in between them. We reformulate the law of social gravity, where the ratio between attraction and repulsive forces is related with the ratios of degree centralities of one social actor and the all other social actors. If the social activity of a social actor can compete with the social activities accumulated by the whole community, the attraction force increases (i.e. ratio gets closer to 1), namely social actor get closer to the centre. Otherwise it becomes smaller and disappears as the attraction force decreases.

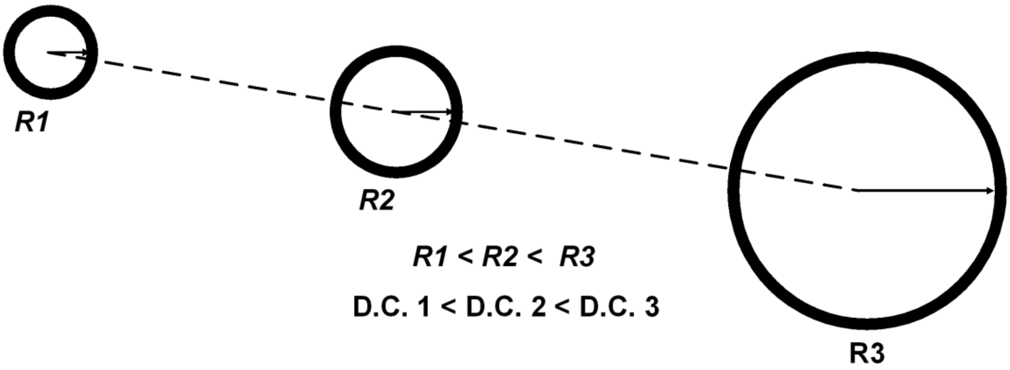


Fig 6.1 nodes placed according to their degree centralities

6.1 Pressure model for non-temporal datasets

Given a list of nodes; L = $\{n_0,....n_n\}$ and list of edges E = $\{e_0,....e_m\}$ P_{Space} is calculated by:

$$P_{Space} = \frac{\sum_{i=0}^{i=g} n_i}{V_{Space}} \qquad (28)$$

n_i is the degree centrality measure of the i^{th} node in the graph, g is the number of nodes, and $V_{Space}$is the absolute volume of the space, which can be set by the user. Calculation of n_i is given by:

$$n_{\mathrm{i}} = \frac{\sum \mathrm{l}}{\mathrm{g}-1} \qquad (29)$$

l is the total number of direct relations of i$^{\text{th}}$ node with other nodes; and g is the total number of nodes in the graph. According to our IGL formulation (1), if we omit parameter T and R namely if we use isothermal formulation of IGL, the volume of a gas will given by:

$$V = \frac{n}{P_{Space}} \qquad (30)$$

And we use this formulation to obtain volume of a node *V*, where *n* refers *the degree centrality* namely n_{i} of a node.

With this formulation, we guarantee distributing the space volume $V_{Container}$ to all nodes directly proportional to their degree centralities. A size increment of a node will result in the size decrement of others according to their degree centrality values, which is actually an automated visual clutter reduction mechanism that we propose in sub-section 6.4.

6.2 Pressure model for temporal datasets

Temporal social network analysis data provide users to understand communication structure of a social group [22]. We descritize continues time with a time slice value Δ*t*, which is decided by the user, to establish a temporal basis. In time domain, computation of space pressure P_{Space_t} at time *t* is performed by:

$$P_{Space_t} = \frac{\sum_{i=0}^{i=t}\left\{\sum_{j=0}^{j=g} n_{it}\right\}}{V_{Space}} \qquad (31)$$

g is the number of nodes in the graph, n_{it} is the accumulated degree centrality until time *t* is given by:

$$n_{i_t} = \sum_{i=0}^{i=t}\left\{\frac{\sum 1}{\mathrm{g}-1}\right\} \qquad (32)$$

Moreover, computation of the volume of the i$^{\text{th}}$ node at time *t* is done by:

$$V_{i_t} = \frac{n_{i_t}}{P_{Space_t}} \qquad (33)$$

With these formulations accumulation of temporal data is done by summing up all centrality measures up to time *t*, as a result positions of nodes become time dependent, all established relations in the space affect the layout.

6.3 Layout

Large data representation is difficult, and applications of focus-context techniques as the hyperbolic views are good solutions for this problem. In Euclidean space, the circumference of a circle increases polynomialy as its radius increases, but in hyperbolic space, circumference of a circle increases exponentially. This property yields more room to use, which supports focus-context views [11]. For more descriptive information about hyperbolic layouts we refer to [11].

Our technique attempts to place similar nodes close to each other [8]. For non-temporal datasets, our similarity metric is the degree centrality value. All nodes are stored in an array, and sorted according to their degree centrality values. In temporal graphs, similarity of a node is given by the degree centrality value and initial time value. For every time slice Δt, we hold a node array $\Delta t_N = \{n_1 \ldots n_n\}$, and sort according to their degree centrality values. If a node becomes visible at time Δt+1, it is added to corresponding array and participates in sorting procedure.

Sorted nodes are placed on circular orbits. The number of orbits is calculated in accordance with the number of nodes to distribute nodes evenly. If required amount of nodes are placed on the current orbit; next node is placed on the next orbit.

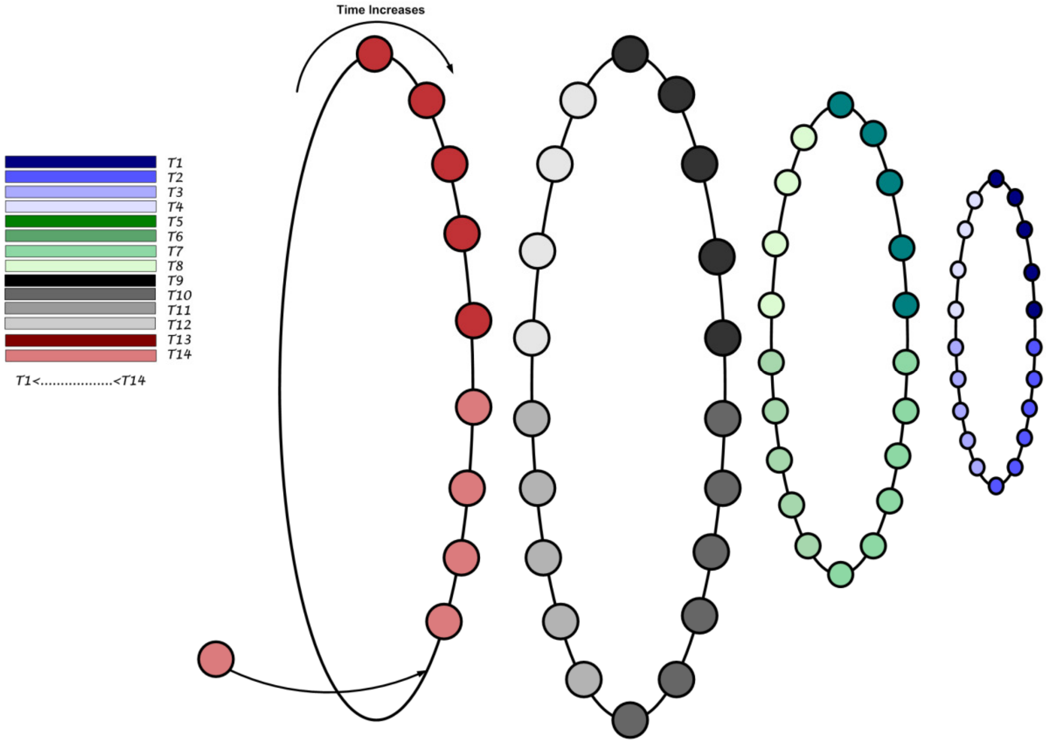


Fig 6.2 nodes are placed on circular orbits as time passes

6.4 Visual Clutter Reduction

The users of information visualization software should not loose focus of attention [8, 9]. However, when interacting with complex datasets, it is challenging to maintain focus due to human cognition system limitations [4, 6]. Therefore it is necessary to perform some kind of visual abstraction [33, 34]. The proposed technique makes use of IGL to transform less active nodes in size and position without modifying the layout hence the perception of the user is not disturbed. Moreover according to experiments conducted by Kadaba et al [45] such transformations improve comprehension of complex casual relations.

The proposed representation and layout technique implies that under certain pressure and volume conditions only a limited number of nodes are rendered on screen.

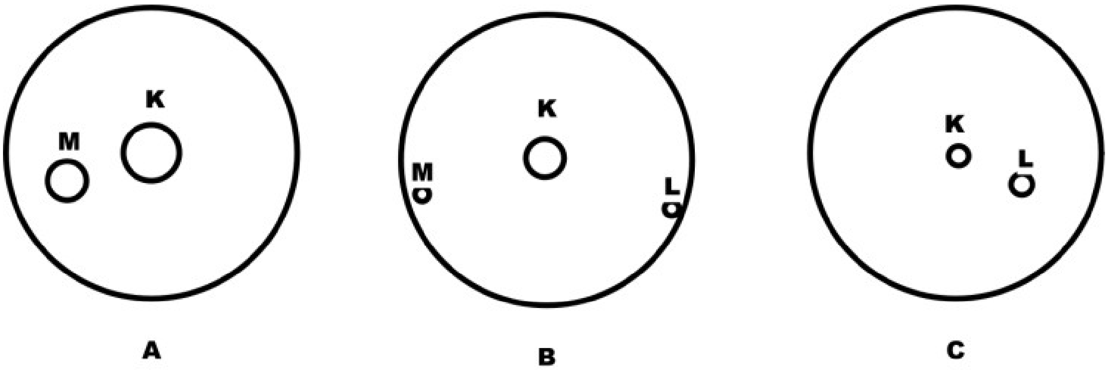


Fig. 6.3. Automatic visual clutter reduction due to increasing space pressure over time. Actor M gets smaller at time value B and then disappears at time value C. Actor L gets larger and actor K remains at canter zone, but gets smaller due to increased pressure.

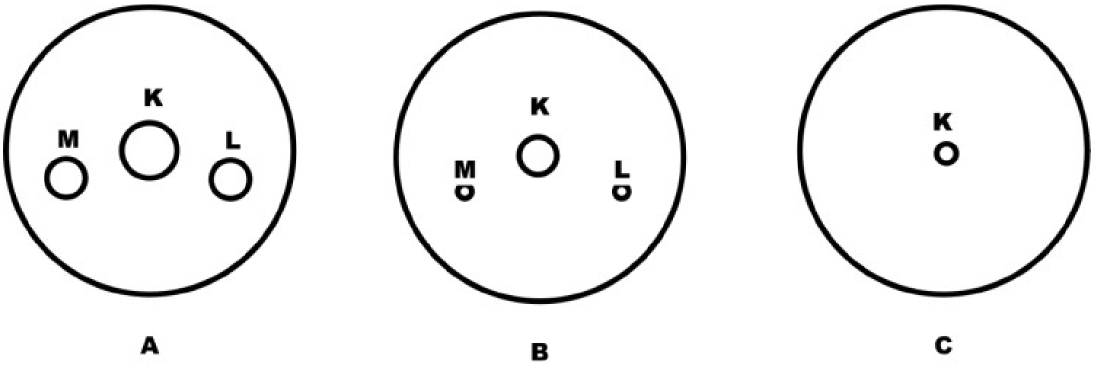


Fig.6.4 Users can control the visual clutter reduction. In this example the time value does not change, but the space volume is decreased by user at B and C. Thus the nodes get smaller. Note that nodes keep their positions, because the volume ratio between users remains constant.

The space pressure is related with the degree centrality of all actors and increases with time. In this case nodes with small degree centralities get smaller and translated far away from the centre zone. Eventually some of them become invisible. This is an automatic visual clutter reduction method which does not give the user any control over the process.

As reported by Eick [46] and Ellis [47] it is important to give users control over visual clutter reduction. We have added functionality to graphical user interface for users to let them increase and decrease the space volume. This feature allows users to control the number of nodes rendered on screen.

With the proposed visual clutter mechanism we obtain a visual representation that has the enough information density: *neither too cluttered nor too sparse* [59].

6.5 Visual Encoding

Implemented visual encoding schema applies nodes as signifiers for social actors. As the importance level of a node changes, it size enlarges or shrinks, following the IGL analogy. Besides, the saturation alters as importance level of node alters. As a result important nodes have larger sizes with high saturation. The colour of the node becomes red only when an anomaly is detected. Combination of this schema with our layout algorithm, in which important nodes are drawn at the very canter of the viewport, we can obtain self descriptive images.

Representation of edges is controlled by the user in runtime. Colour of an edge is interpolated between red and green, where green side is a visual indicator of the source and red side is used to indicate the destination. Another important advantage of our tool is the minimizing disorientation of the user. In some focus context visualization systems users can lose their orientation while they are exploring the dataset and we tried to minimize this by representing the whole layout in one image, which increases the information density is increased. Our representation method can be compared with existing hyperbolic layouts.

Table 1 Comparison on the information density with existing systems

Visualization Method	Number of Visible Nodes	Visibility of the Layout
H3 (and others like Walrus)	<2,000	0<# nodes< 5.000
2D Hyperbolic Tree Viewer	Hundreds	0 < # nodes < 478
IGL	>50,000	0< #nodes< 50.000

7 Anomaly detection

Anomaly detection is a classification problem where an act is classified as *normal* or *anomaly*. Anomaly detection is an important aspect of sociology and related domains. The aim of detecting an anomaly is to decrease or, if possible, prevent the amount of damage by applying necessary strategies [24]. In sociology, approaches for detecting anomaly are mostly based on statistics. Statistics is a field of mathematics aims to analyze, and interpret the data [50]. As described in [51, 52] anomaly detection can be summarized by four tasks.

1) The statistical model of the *normal behaviour* must be obtained.
2) Every new act is compared with the available model
3) The model must be updated.
4) If the 'new act' does not overlap with the statistical model the necessary indicator for an anomaly must be produced.

Self organizing map (SOM) is an unsupervised learning method [53, 54] is widely used to detect anomalies where every social actor (nodes) modelled with a weight vector. At initial step SOM builds the weight vectors based on sample dataset to build the normal models, which will be used to detect anomalies and learn new trends. Because of the initial task, SOM is not an efficient method for temporal inconsistent systems like Enron Corpus where the number of nodes alters at any time which requires the repetition of the initial step when new node added.

Linear Regression analysis is another approach to detect anomalies through composing linear line which 'best fits' the datasets on an Euclidean coordinate system. It is not trivial to compose the regression line and there are methods like *least squares* to produce the points of the linear line. In linear regression analysis anomaly is decided by comparing the Euclidean distance of the corresponding point of the 'new act' with the estimated error [55, 56]. As the comparison of the 'new act' done through with a **linear line** with slope m, linear regression analysis does not suitable for non-linear systems. [55, 56] As in the figure:

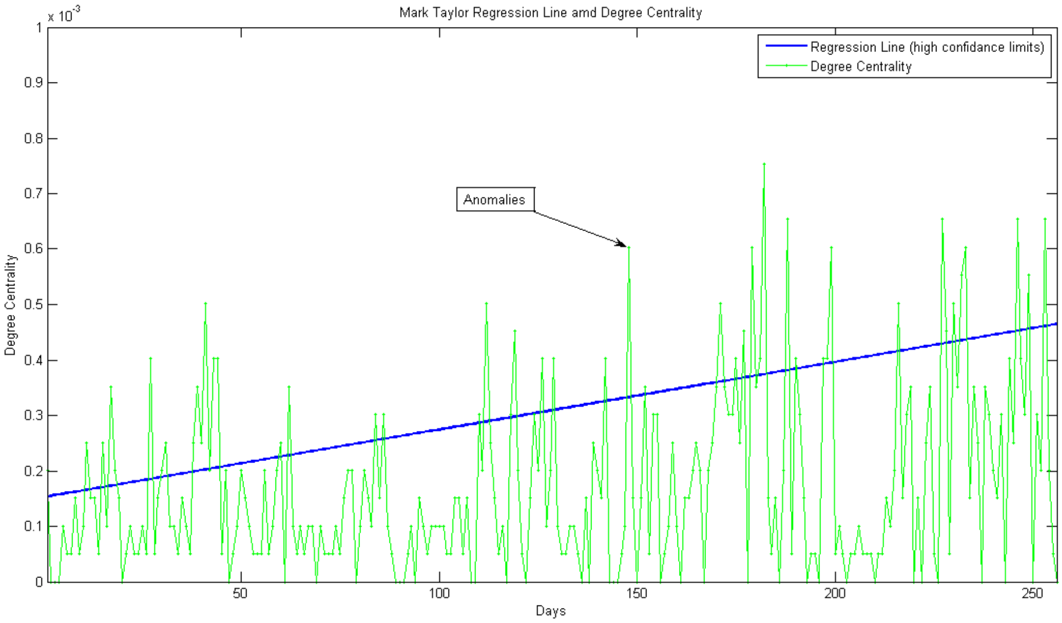


Fig 7.1 Degree centrality versus Regression line for Mark Taylor

We have developed two distinct methods to detect large anomalies for non-linear systems. The proposed anomaly detections are adaptive where all social actors have their own discrete normal model that is evolving as new acts produced, and addition or deletion of an social actor do not disturb anomaly detection model of the other social actors. Moreover, the proposed methods applies unsupervised learning scheme as the normal behaviour of a social actor is decided by the all social activities happened until a given time.

The first method detects the change of communication density of one actor. The second one alarms on the change of number of communication partners of an actor.

7.1 Density Based Anomaly Detection

The first technique is based on the correlation between increment rate of degree centrality and increment rate of social activity value (SAV). SAV is obtained by formula 1 where this time *l* refers to *the number of all communications* of social actor *i* and *g* is the *all performed social activities belongs to the social actor i.*

In a social structure one’s degree centrality may increase rapidly with unimportant activities like spam e-mails; however, if the increment rate of SAV measure (i.e. messaging with an old friend who was absent for a long time) is equally steep at a given time, it can be an indication of an anomaly. To detect it, change rates of degree centrality and SAV are stored separately, and their average values in a time window (0,..,j) are calculated. The two change rate summations (CRS) for a node *i* is given by:

$$CRS_{DC_i}(t) = \left|\sum_{i=0}^{i=t+1}\{n_{i_{t+1}}\} - \sum_{i=0}^{i=t}\{n_{i_t}\}\right| \qquad (34)$$

$$CRS_{SAV_i}(t) = \left|\sum_{i=0}^{i=t+1}\{m_{i_{t+1}}\} - \sum_{i=0}^{i=t}\{m_{i_t}\}\right| \qquad (35)$$

where *m* is the SAV value of a node. The average values $DC_{avrg}$and SAV_{avrg} are used as anomaly thresholds. The thresholds are given by:

$$DC_{avrg} = CRS_{DC_{\mathrm{i}}}(t)/t \qquad (36)$$

$$SAV_{avrg} = CRS_{SAV_{\mathrm{i}}}(t)/t \qquad (37)$$

If both the centrality value and SAV of a node n at time *t* is higher than the threshold values, [$\mathrm{n}_{\mathrm{DC}} > DC_{avrg}$] ∩ [$\mathrm{n}_{\mathrm{SAV}} > SAV_{avrg}$], we mark them as anomalous actors.

7.2 Frequency Based Anomaly Detection

The second technique is observing important alternations of the degree centrality value together with its occurrence in a given time window. This type of anomaly detection observes unusual big alternations of the increment rate of degree centrality. If change rate (CR) of degree centrality of node i is higher than the average degree centrality and the frequency of this rate in the time window is low, it is detected as anomaly and is given by:

$$CR_{DC_i}(t) = |n_{i_{t+1}} - n_{i_t}| \tag{38}$$

$$freq_{\mathrm{VAL}} = freq\left(avrg < CR_{DC_{\mathrm{i}}}\right) \tag{39}$$

If $\mathrm{freq}_{\mathrm{VAL}}$<2% this is an evidence for anomaly.

The proposed methods can detect anomalies of

1) Individuals
2) Group of people
3) The whole social structure.

As decided by the user. Besides, user can switch the anomaly detection method to see different type of anomalies.

8 Case Studies

In this section we show visual representations of datasets that applied to validate and evaluate our method. In first section we represent the results of synthetic datasets. In the second section we show visual representations of real-life datasets.

8.1 Synthetic Datasets

We generated simple datasets, which practice special type of relations, and compare visual representations with our expectations. Each dataset contains 100 nodes with 30.000 edges. All datasets are temporal and cover a specific type of relations

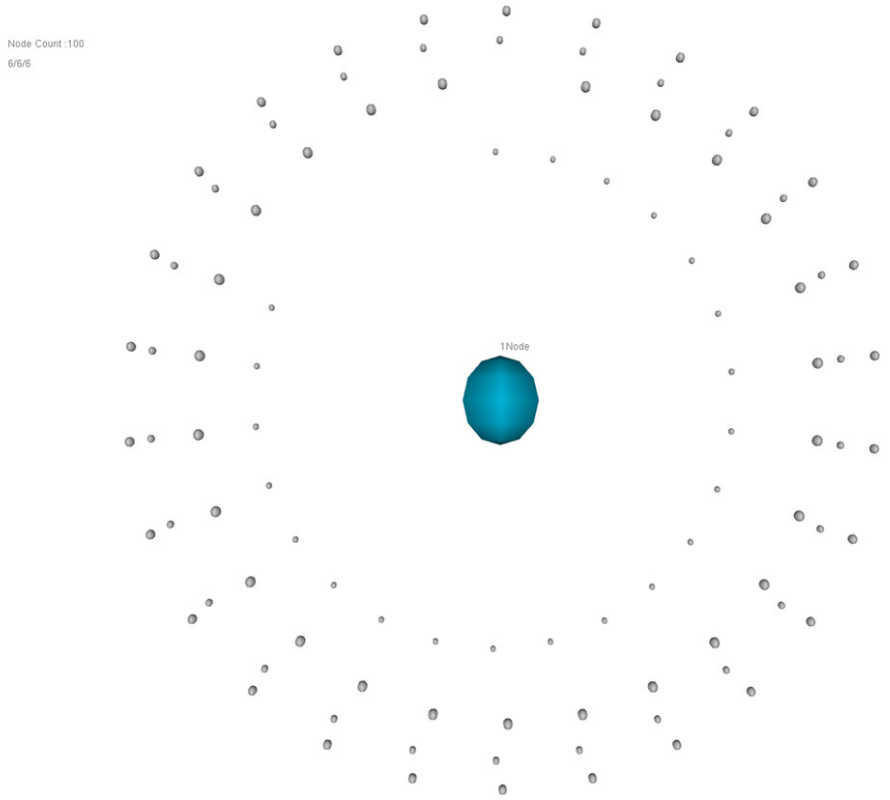


Fig 8.1 Many to one

Many to one relation: This refers to a relational pattern, in which all members in a social group have only direct link with a single unique actor. This actor is the center of social network; every actor can communicate with each other only through this unique actor. In this scenario the central actor is represented with a large balloon due to its high degree centrality and located at the center of the unit sphere. Others dispersed around this actor at equal distance.

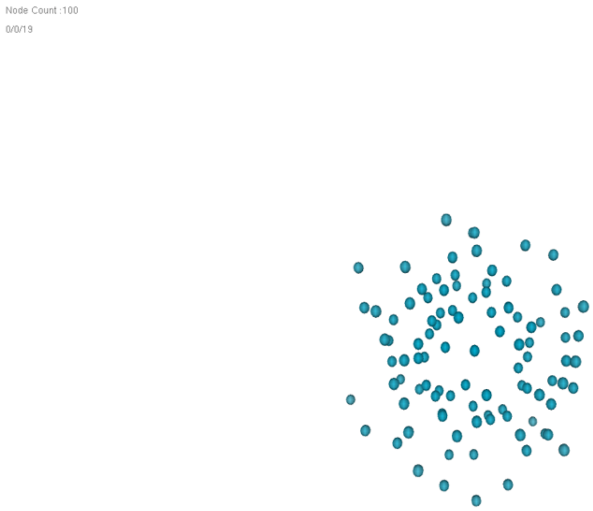
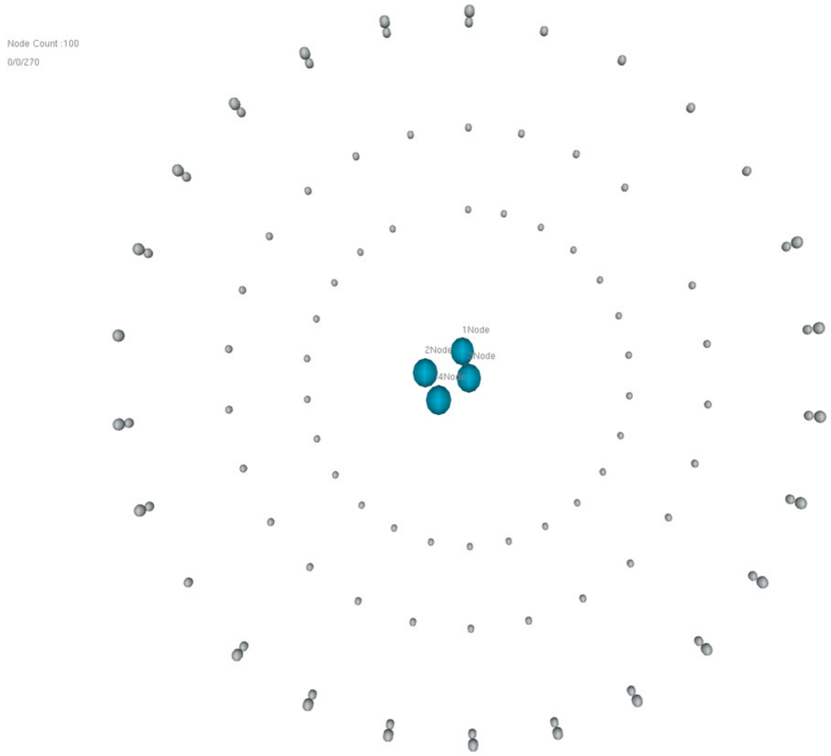


Fig 8.2 Many to Many

Many to many relations: This is a social structure, in which every members have direct links to all members. It is not centralized. Actors can communicate and share information with each other through direct links. This network produces a centralized topology, where all actors have similar sizes and are clustered near the center.



8.3 Many to some

Many to some relations: Many to some relations occur, where all members in a social group have direct links with a small number of actors. This framework forms a central group, which has direct links with all other actors in the social group. A representation of this social structure reveals the central group by positioning them at the central area of the hyperbolic space. Other nodes are positioned around this group.

At all three experiments the system performs inline with our estimations. The results are highlighting the important actor(s) of the social network correctly and visually appealing.

8.2 Real Life Datasets

8.2.1 Static Datasets

In this section we present the outputs that we obtained from SocialNet dataset. SocialNet is a dataset of social interactions (180 direct relations) of 129 individuals. As our framework represents important nodes at the central zone of the space via calculating degree centrality values, the most important actors for SocialNet can be identified.

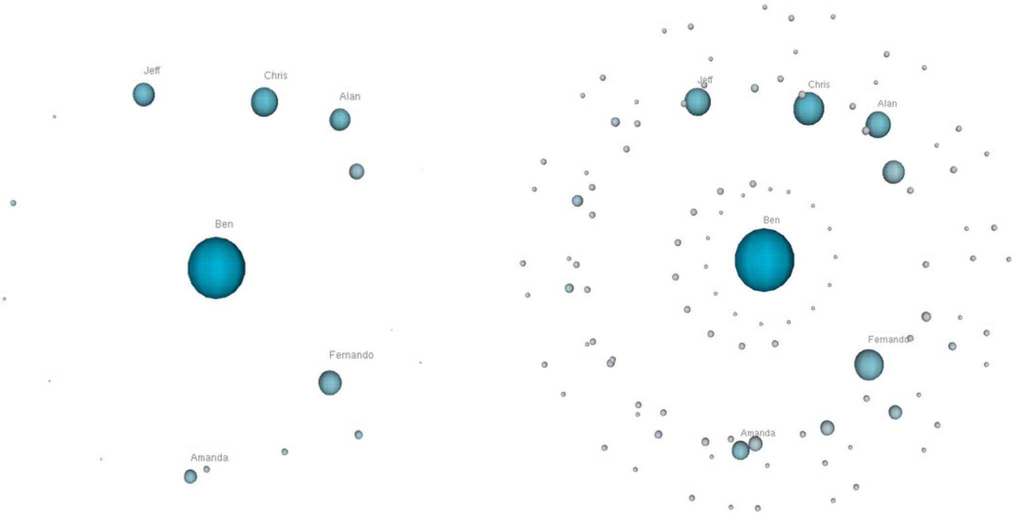


Fig. 8.4. Users can adjust the level of visible nodes by modifying the space pressure. At the left image, pressure of the space is larger than the pressure at the right image. In both cases "Ben" is at the center as the most important social actor in SocialNet. He is followed by "Cris", "Jeff", "Alan", "Fernando" and "Non", respectively.

8.2.2 Temporal Datasets

In this section, we discuss representations of temporal networks. We present features of our technique on Enron and Newsgroup dataset.

8.2.2.1 Enron Dataset

The Enron e-mail dataset is a real-life temporal social network data of a large organization which is available online at CMU web-pages. Original data contains

619.449 emails belongs to 158 Enron employees. For detailed investigations of this dataset we refer to [29, 30]. The importance of the Enron e-mail dataset is two-fold: it is a dataset belongs to real-life organization, and the corresponding organization collapsed after a crisis.

8.2.2.1.1 Data Pre-Processing

The data is distributed to distinct folders according to the names of Enron employees including CEO's and executives and every folder has it's sub-folders like Sent, and Inbox. Each e-mail has the sender-receiver, date, subject and the context data. However, some e-mails contain missing information like receiver address' or the date of the e-mail is before 1980 so we cleaned the dataset through neglecting the followings from the dataset.

1) E-mail postings earlier than 5/01/1998
2) E-mails with no recipient(s) / Sender information
3) E-mails with same sender and receiver address.

After the clean process we have obtained 306.133 e-mails belongs to 19.898 different e-mail addresses accumulated in between 5/01/1998 - 12/07/2002.

8.2.2.1.2 Visualizations

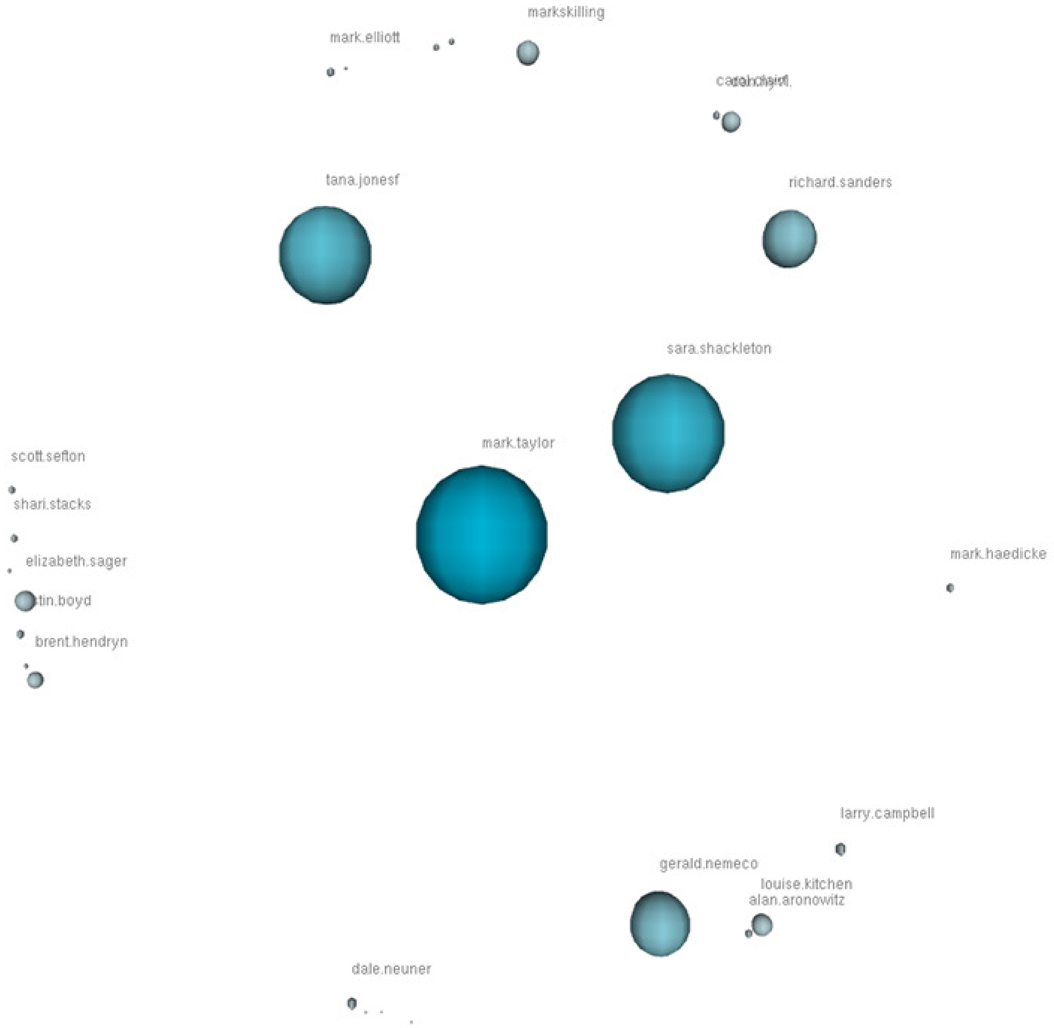


Fig.8.5. Snapshot of Enron data on 14.9.1999 with 1016 actors. We see that “Mark Taylor's” centrality measure is large and the system puts him at the canter of the space sphere. We can identify four other highly active actors followed by nine others.

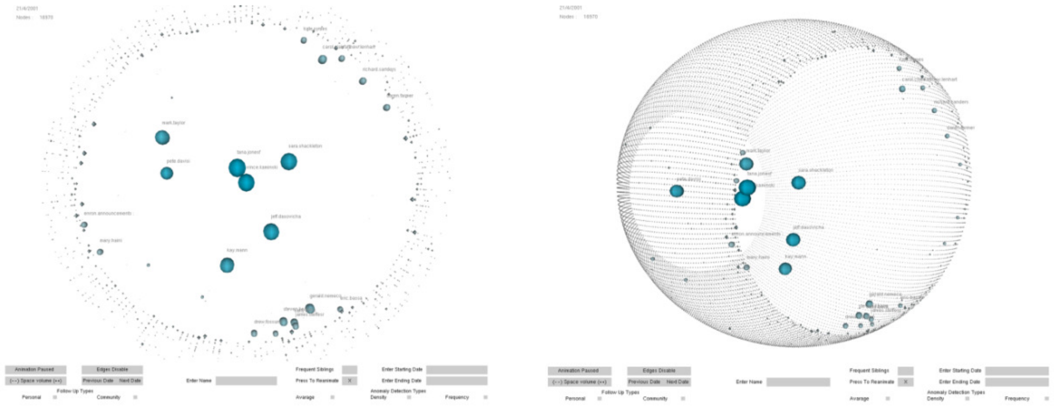


Fig. 8.6. Snapshots of Enron data on 12.7.2001 with 18.839 actors. There are seven highly active actors. Users can increase outer space pressure and fade out unimportant nodes from the scene to improve visual clarity (left). They can monitor overall activity by decreasing the space pressure (right).

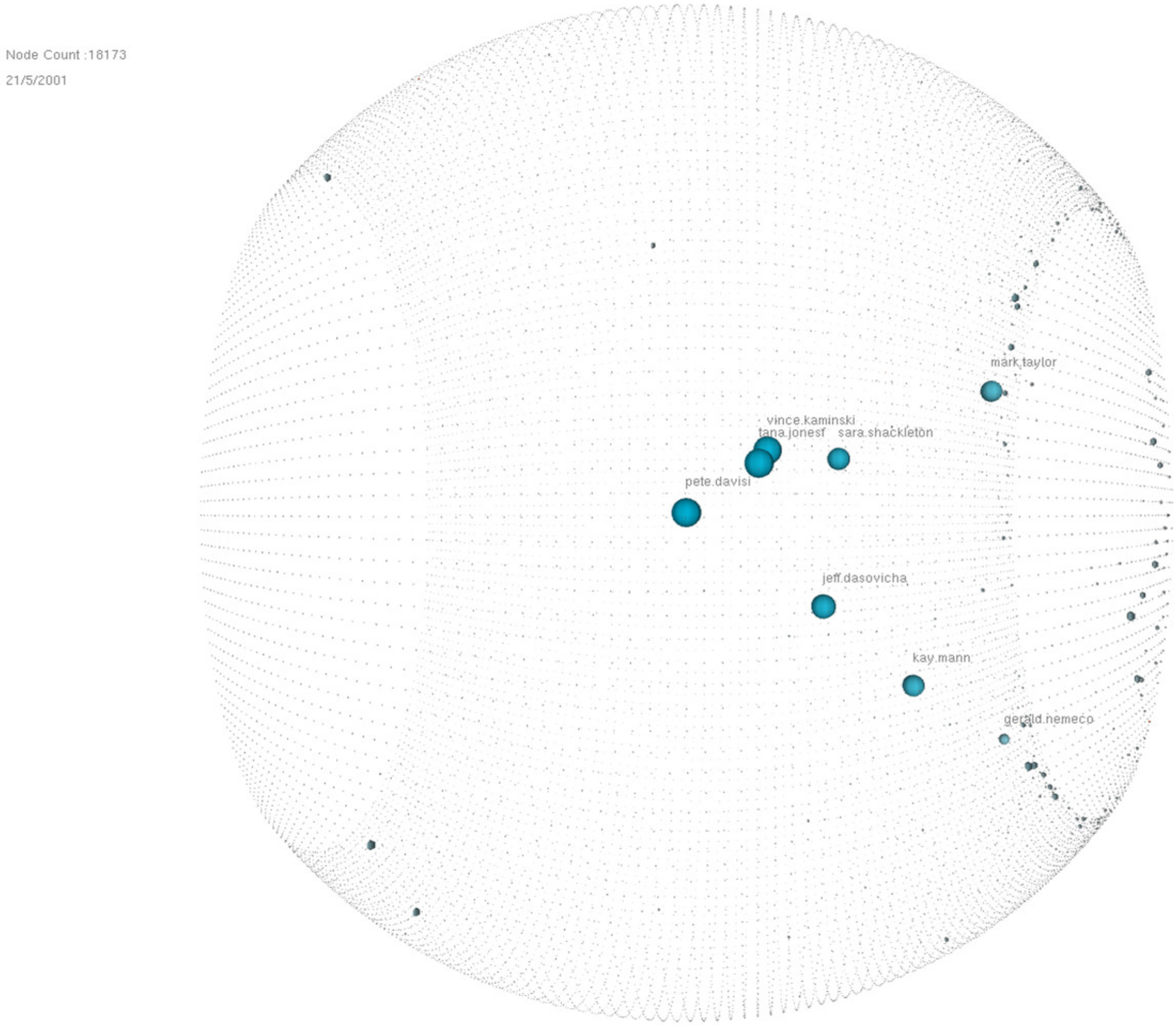


Fig8.7 Enron Corpus at 21/05/2001, 18,173 nodes are visible.

8.2.2.1.3 Anomaly detection

Visual representation of detected anomalies together with the data is a way of increasing the information density of an image and hence is hot topic for the information visualization agenda. As discussed in section seven, several anomaly detection algorithms were applied. Visual representations of these anomalies are as the followings;

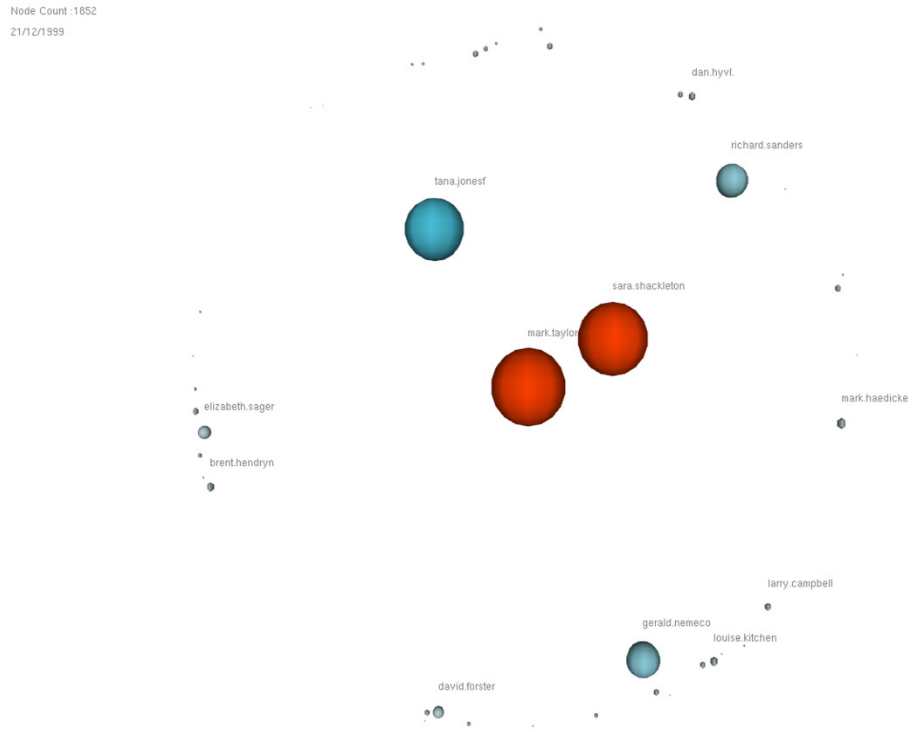


Fig. 8.8. A snapshot of Enron data on 21.12.1999 with 1852 actors. User performs anomaly detection based on communication density change.

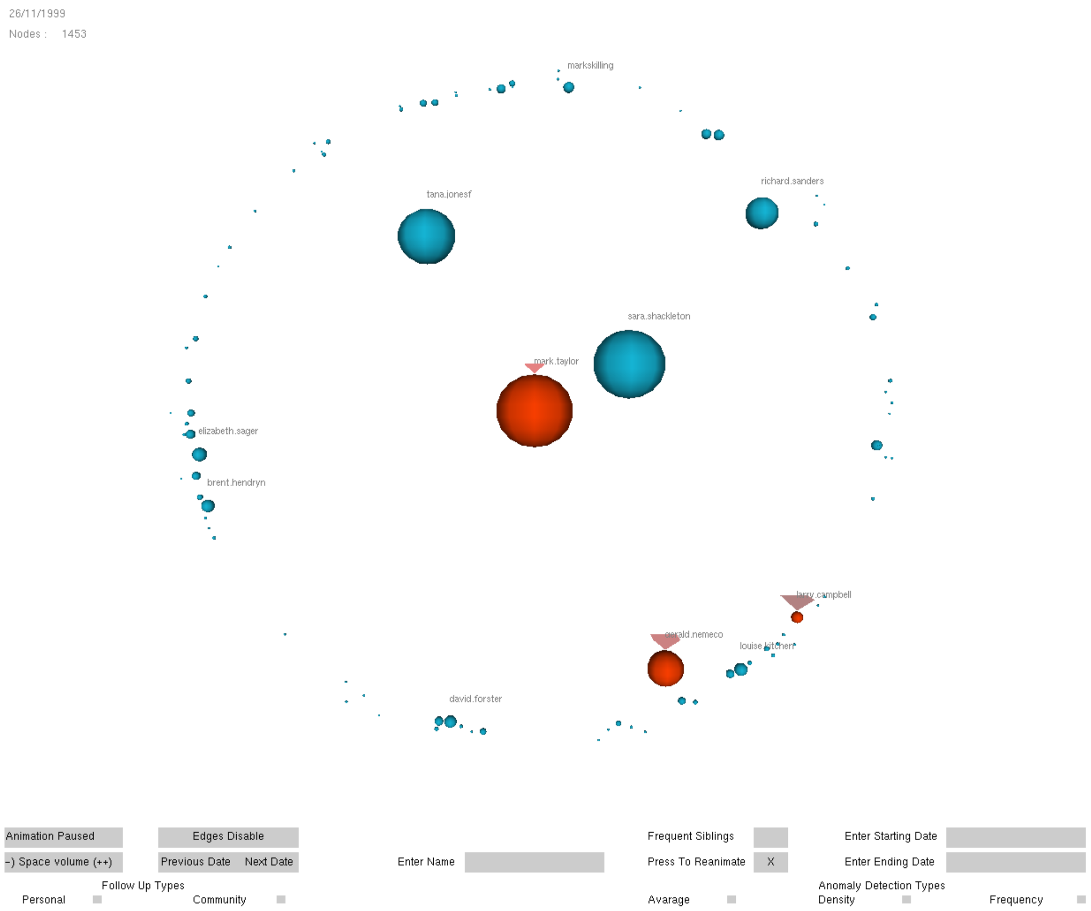


Fig 8.9 detected anomaly of Mark Taylor, Gerald Nemeco and Larry Campbell

8.2.2.2 Newsgroup dataset

This data contains approximately 44.797 postings, which belongs to 20 different newsgroups. These postings sent from 5417 distinct e-mail addresses in one month. With our method, we wanted to see which e-mail addresses are more active during this time.

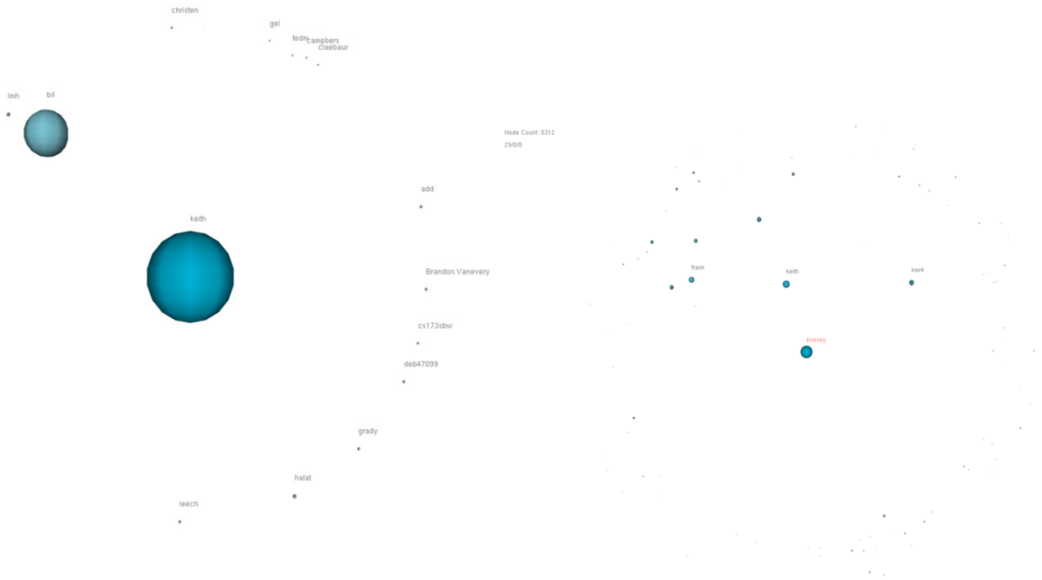


Fig. 8.10. Left: During first weeks of the month, "keith" is the most active actor. Following "keith" "livesey" is the second most active actor. Right: During last weeks of the month, we see that "livesey" becomes the most active e-mail address by pushing "keith" to second. Note that the space pressure gets larger over time and nodes with little degree centrality increase get smaller and eventually disappear.

9 Implementation

The design of the proposed visualization system has two main parts data process and visualization which are implemented with C++ programming language using Glut libraries on Intel Dual Xeon 3.4 GHz with 4 GB of memory. The first step is fetch; the system gets dataset in predefined format and pre-process' the data to compute degree centrality, node volume and complete hyperbolic layout for each time step. The pre-processing should be done only once for each dataset. Because many datasets are too large to hold them in the system memory, proposed visualization framework uses paging algorithm.

Paging is a memory allocation method that loads fixed length of pages into system according to a predefined method. In the proposed system Anticipating Paging is applied where the numbers of pages are decided by the user and fixed number of pages are fetched and deleted as the program runs.

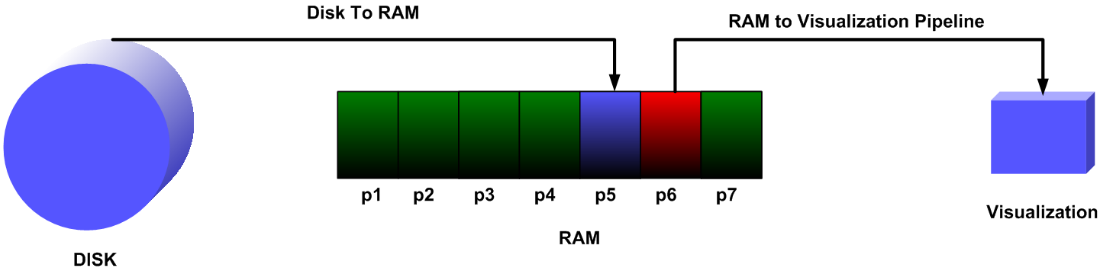


Fig 9.1 Paging procedures

The proposed drawing method is linear with the visible number of nodes. [11]. Because of the IGL formulation, a node is rendered if and only if:

1) SNA metric of nodes value is higher than the average of the summations of all SNA metrics of the social space until a given time.

2) SNA metric of nodes are higher than a user defined threshold.

The two cases work well for the most of the datasets tested but it is fail in one case. If all the social actors of a very large (Over a millions of social actors) social system have same importance, social space will not be rendered. However the described social system is not available in human social life. [27]

As a result the visual and computation complexity of the method is bounded to the number of the visible nodes.

Current version of the proof of the concept program follows the following procedure:

```
ComposeArrays (nodeArray, edgeArray) returns set of node arrays

var Arrays[time_value] ; holds nodes for time ∆tN
var Nodes[nodeArray_Size] ; holds nodes

for timeStamp ← 0 to time_value
    for each edge ∈ edgeArray
        if edge.time <= timeStamp
            source ← edge.source;    target ← edge.target
            if Nodes_Size != 0
                for each node ∈ nodeArray
                    if {source|target} == node
                        increment Degree Centrality({source|target})
                    else
                        insert  (Nodes, {source|target})
            else
                insert  (Nodes, {source|target})
        else
            insert (Sort Arrays[timeStamp],Nodes)

return Arrays
```

9.1 Time Analysis

The system except datasets with 1) no any adjacency information, 2) ordered according to temporal in ascending order. Finding adjacency information of a node is in order $O|EV|$ where E is for edges and V is for nodes, and if the data set is temporal namely computation of the adjacency information is done in T amount of time slices, the order increases to $O|TEV|$. If the supplied dataset has adjacency information then the complexity of the whole layout will become order $O|V|$.

Apart from the pre-processed attributes of the dataset, at runtime the system performs node transformations, anomaly detection and space volume manipulation at interactive speeds. Currently the system is not optimized for speed and memory usage. We integrated the anomaly detection into rendering pipeline as a result anomaly detection algorithm demands $O|s|$ CPU time where s is the window size of the normal behaviour model.

Computation of the layout of a large graph with 60,028 edges and 50,125 nodes that accumulated in 200 time slices takes more than 1 minute which in average 81secs/200 = 0,405 sec to compute one slice, however as the number of edges and time slices increase the computation time increases for example Enron with 936 time slices takes 12 minutes which results 0.79 sec in average. Moreover, this timing contains the delay of disk write operation for the necessary paging algorithm.

Table 2 Time comparison table. The pre-processing (PT) in second (s) (for one slice) and runtime (RT)in frame per second (fps) values are average values measured during the experiments.

Dataset	PT. (s)	Node	Edge	RT (fps)
Enron	0.76	19.989	323.078	25
NewsGroup	0.35	5417	44.797	30
SocialNet	0.0013	129	180	40

The proposed system has one advantage over existing hyperbolic layout methods. Because the proposed method applies paging algorithm it is not neccesarly to load all the dataset into main memory where the system memory is weak.

Table 3 Comparisons of the memory requirements-data size, among existing hyperbolic layouts.

Visualization Method	Dataset (#Edges)	Memory Requirements
H3 (and others like Walrus)	110.000	Min. 1GB
2D Hyperbolic Tree Viewer	Hundreds	128 MB
IGL	>300,000	128 MB

9.2 Usability Study

We conducted user tests with a group of 15 computer science researchers who has no knowledge on the Enron Case. We gave them a short training of our tool before testing. We asked them to perform three quantitative tasks parallel with the following works [61] to measure the cognitive complexity of our images produced for a large dataset; Enron dataset.

1) Identify four most important actors on 1-1-2001.
2) Count the number of visible nodes on 13-7-2001.
3) Write down the number and the dates of detected anomalies for “Mark Taylor”.

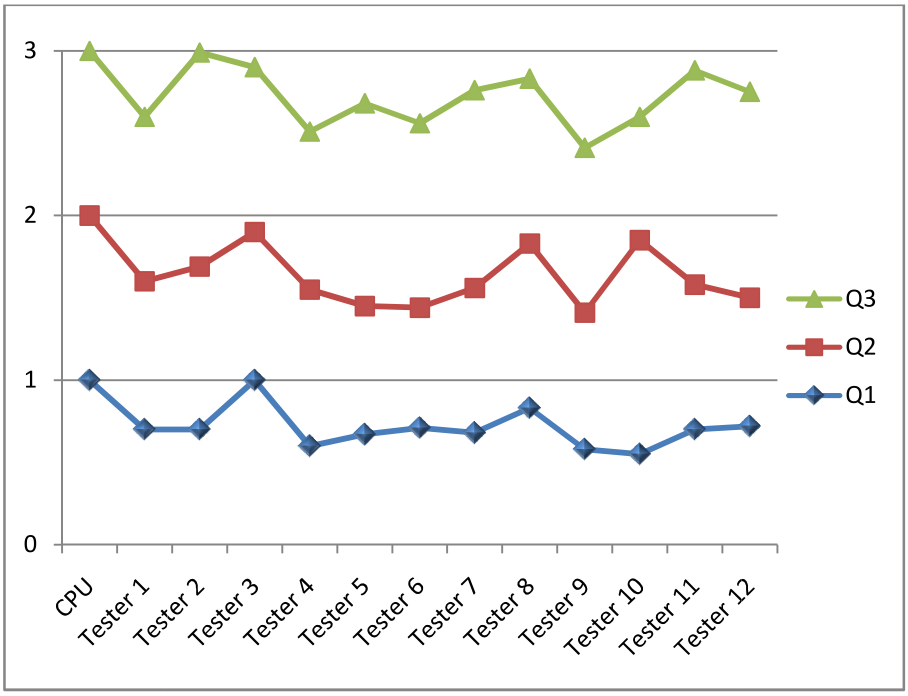


Table 4 usability test results

The participants were asked to perform these tasks in 10 minutes. We have compared their answers with actual values and computed the error rate. Users completed the first task with 71%, the second one with %82 and the third one with 74% accuracy. They expressed that the visualizations are self-explanatory and visually appealing.

During the tests participants asked for more information on several actors and their involvement in Enron case in particular: Mark Taylor, Vince Kaminski and Tana Jones. They also wanted to know why “Pete Davis” appeared all of a sudden and become an important actor.

10 Discussion and Future Work

In this thesis we presented an interactive visualization technique based on IGL and cannon of creativity theory to explore temporal social networks. Our technique uses position, size and colour for visual encoding of information. We have shown that with the presented technique users can perform basic tasks on complex social network datasets. Users can monitor large (N>50.000) social network datasets with more than 30 updates per second. IGL analogy allows users two distinct techniques to reduce visual clutter caused by high number of nodes. Currently our system can handle data more than 50.000 nodes at interactive speeds with generated test data.

We provide anomaly detection based on communication density and communication partner number. Moreover users can select and track particular actor and investigate its five most active communication partners over time in detail.

Main strength of the proposed technique can be summarized as follows;

1) Properties of large temporal social network datasets can be explored.
2) As the importance calculation of social structure concerns past data, any visual depiction summarizes all social actions.
3) The new pressure model provides easy to implement visual clutter and layout method.
4) The proposed tool is not only for Temporal Social Network Datasets but also any type of temporal dataset, which elements can be ordered. For instance, to represent evolution of the popular genre in the movie industry for past 10 years.

Apart from the strengths the proposed technique has several drawbacks.

1) It is not able to show a graph

2) Pre-processing limits the interaction

There are several results obtained from this work as well. The first important result is the importance of Geometry for applied computer graphics. It is clear that hyperbolic space provide us to represent larger datasets, so any relevant updates about hyperbolic geometry or any manifold must be observed and applied. This work also shows that still there are available physical abstractions that can be applied to social studies to find properties of, and represent social structures.

In future other social network analysis metrics will be implemented and paging algorithm will be optimized. Moreover, the proposed method will be implemented on GPU to decrease the work load of the CPU so, implementation of advanced interaction techniques become possible. In addition the proposed technique will be evolving to visualize large temporal heterogeneous datasets.

Appendices I

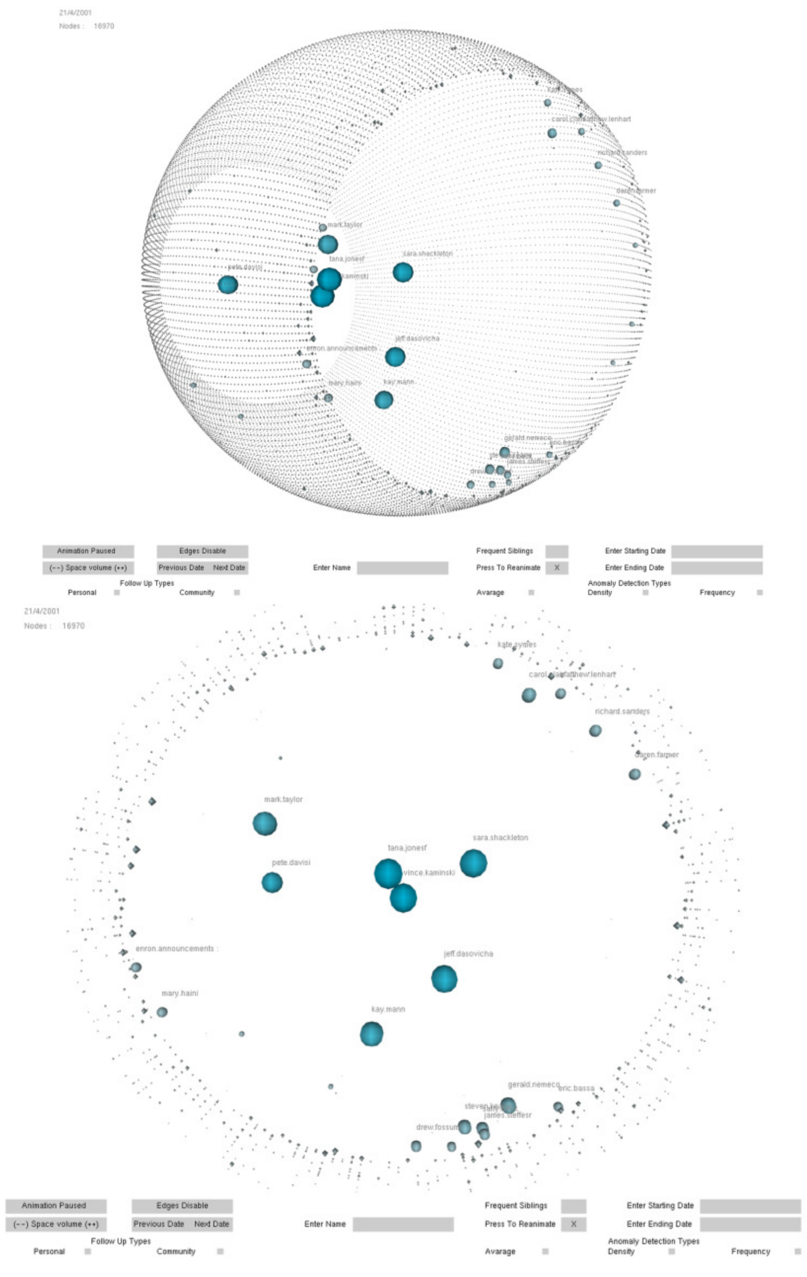
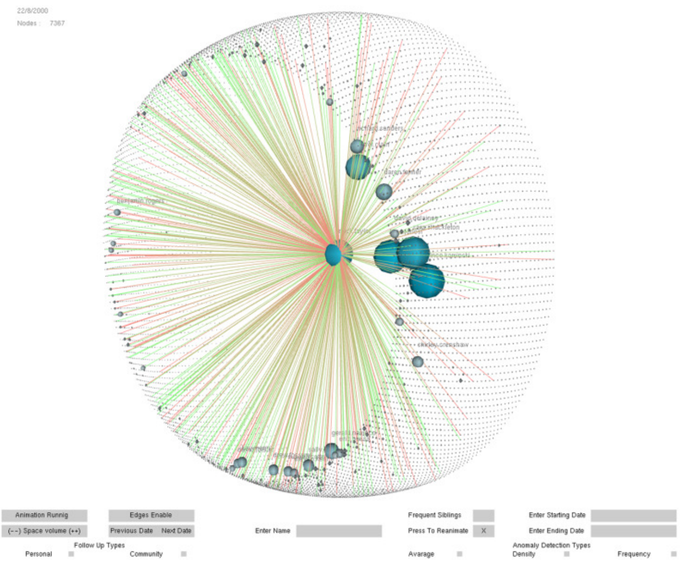
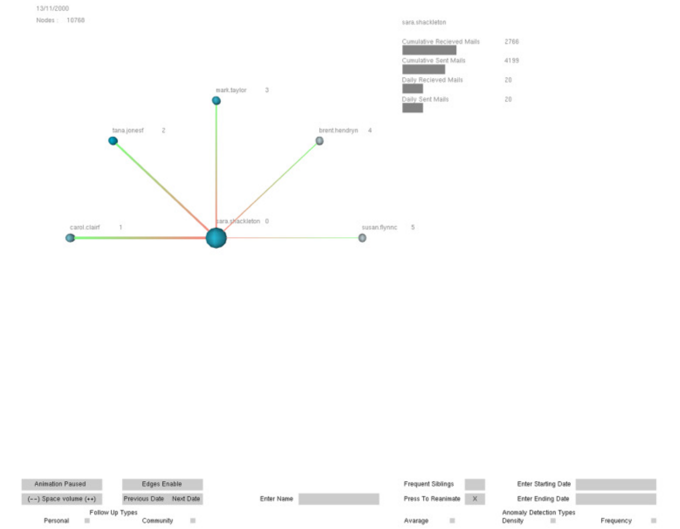


Fig. A.1 Screen captures of Enron data on 24.1.2001 with 16.970 actors with user interface. On the upper left there is the date and number of active nodes. On the bottom you can see the graphical user interface consist of search, animation and abstraction control buttons. The upper and lower images are from the same dataset with same time value but different space volumes. The user manupilate volume through the UI "space volume" button on left bottom corner.



22/8/2000

Nodes : 7367



13/11/2000

Nodes : 10768

sara.shackleton

Cumulative Recieved Mails 2766

Cumulative Sent Mails 4199

Daily Recieved Mails 20

Daily Sent Mails 20

Fig A.2 Screen captures of Enron data on 22.8.2000 with 7367 nodes. The left image shows incoming (green) and outgoing (red) email traffic of one active actor. Users can search a user by name or pick during time series animation to switch to an investigation screen (right image). This screen depicts five most active communication partners of the selected actor and prints communication load data as bar graph on top right corner of the screen.

References

[1] Kamada T. and Kawai S. An algorithm for drawing general undirected graphs, Inform Process pp:7-15, 1989

[2] P. Eades “A heuristic for Graph Drawing” Congrassus Numerantium 42(1984), 149 – 160

[3] Thomas M. J. Fruchterman, Edward M. Eingold Graph Drawing by Force Directed Placement Software—Practice and Experience, Vol. 21(1 1), 1129-1164, 1991

[4] U.Brandes, Drawing on Physical anologies, Drawing Graphs: Methods and Models, pg. 71-86

[5] S. G. Kobourov, K. Wampler. Non-Euclidean Spring Embedder, *IEEE, transactions on Visualization and Computer Graphics*, Vol. 11, No. 6 pp 757-767, 2005

[6] I. F. Cruz, R. Tamassia “Graph Drawing Tutorial”, Worcester Polytechnic Institute, Brown University,

[7] I. Herman, G. Malençon, Graph Visualization and Navigation in Information Visualization: A Survey. *IEEE, Transactions on Visualization and Computer Graphics*, 2000

[8] H.C. Purchase, Metrics For Graph Drawing Aesthetics, Journal of Visual of Visual Languages & Computing, 501-516, 2002

[9] Di Batista, P. Eades, Algorithms for Drawing Graphs: an Annotated Bibliography. 1994

[10] M. Phillips, C, Gunn, Visualizing Hyperbolic Space: Unusual Uses of 4*4 Matrices The National Science and Technology Research Center for Computation and Visualization of Geometric Structures, 1991

[11] T. Munzner, Lying out large graphs in Hyperbolic Space. *IEEE, transactions on Visualization and Computer Graphics* 1997

[12] J. W. Canon, W. J. Floyd, R. Kenyon, W. R. Parr, Hyperbolic Geometry, Flavors of Geometry, MSRI 1997

[13] G. Leibon, Scissors Congruence: The Birth of Hyperbolic Volume, Department of Mathematics, Dartmouth College.

[14] J. Walter, J. Ontrup, D. Wessling, H. Ritter, Interactive Visualization and Navigation in Large Data Collections using the Hyperbolic Space, Proceedings of the 3. *IEEE, International Conference on Data Mining* 2003

[15] G. Kumar, Visual Exploration of complex time varying graphs. *IEEE, Transactions on Visualization and Computer Graphics*, 2006

[16] Peter A. Gloor, Capturing team dynamics Through Temporal Social Surfaces, *Proceedings of the Ninth International Conference on Information Visualisation*

[17] Zhumdall, Chemistry 5th edition

[18] W. Müller, Visualization Methods for Time-Dependent Data- An Overview, Winter Simulation Conference,2003

[19] S. Wasserman, K. Faust, Social Network Analysis methods and applications, Cambridge U. Press 1999

[20] J. L. Moreno Who shall Survive? Beacon, 1953.

[21] J. L. Moreno, Foundations of Sociometry : An Introduction, American Sociological Association 1941.

[22] J. L. Moreno, The Three Branches of Sociometry: Apostscript Sociometry Vol.11, No.1/2. 1948 pp121-128

[23] K. Wrightson, English Society 1580-1680, Rutgers University Press, 1984

[24] J. M. McCormick, Evaluating Models of Crisis Behaviour: Some Evidence from the Middle East, *International Studies Quarterly*, Vol: 19, No:1, 1975 pp 17-45.

[25] G.Allison, P. Zellikov, The Essence of Desicion: Exploring the Cuban Missile Crisis, Longman, 1999.

[26] Aage B. Sorensen, Mathmatical Models In Sociology. Department of Sociology, University pf Wisconsin, pp 345-371, 1978

[27] Ronald S. Burt, Models of Network Structure, Department of Sociology, Universtiy of California, Berkeley. Pp 79-141, 1980

[28] F. Harary *Mathematics Magazine*, Vol. 42, No. 3. (May, 1969), pp. 146-148.

[29] J. Diesner, T. L. Frantz, K. M. Carley, Communication Networks from the Enron Email Corpus: “It’s Always About the People. Enron is no Different”, *omputational & Mathematical Organization Theory*, 2005

[30] P.S. Keila, D.B. Skillicorn, Structure in the Enron Email Dataset, School of Computing Queen's University

[31] W. Eberle, L. Holder, Discovering Structural Anomalies in Graph-Based Data. Proceed-ings of the *IEEE Intelligence and Security Informat-ics Conference*, 2006.

[32] C. Chen, An Information-Theoretic View of Visual Analytics. IEEE Computer Graphics And Applications 2008

[33] T.C. Sprenger, R. Brunella, M.H. Gross, H-BLOL: A Hierarhical Visual Clustering Method Using Implicit Surfaces. *IEEE, transactions on Visualization and Computer Graphics* 2001.

[34] M. Balzer, O. Deussen, Level-Of-Detail Visualization of Clustered Graph Layouts Asia-Pacific Symposium of on Visualization, 2007.

[35] J. Heer, S.K. Card, J.A. Landay, prefuse: a toolkit for interactive information visualization, Interactive Information Visualization, CHI 2005.

[36] N. Sawa, A Multiple-Focus Graph Browsing Technique Using Heat Models and Force Directed Method. *IEEE, transactions on Visualization and Computer Graphics,* 2001

[37] E. R. Gansner, Y. Koren, Topological Fish-eye view for Visualizing large graphs, *IEEE, transactions on Visualization and Computer Graphics*, Vol 11 No. 4, pp 457-468, 2005

[38] A. Dekker, Visualization of Social Network using CAVALIER, ACM, Asia Pasific symposium on Information Visualization , 2001

[39] N. Henry, J. Fekete, M. J. McGuffin, NodeTrix: A Hybrid Visualization of Social Networks, *IEEE, transactions on Visualization and Computer Graphics*, 2007

[40] J. Moody, D. McFarland, B. deMoll, S. Dynamic Network Visualization: Methods for Meaning with Longitudinal Network Movies. 2004.

[41] V. Batagelj, A. Mrvar, Pajek "Program for Analysis and Visualization of Large Networks User Manual" version 1.22

[42] http://people.csail.mit.edu/jrennie/20Newsgroups/

[43] Mutton, P. Inferring and Visualizing Social Networks on Internet Relay Chat. *InfoVis* Austin, TX, 2004

[44] A. Quigley, P. Eades, FADE: Graph drawing, clustering and visual abstraction Proceedings of the 8th International Symposium on Graph Drawing, 2000

[45] N.R. Kadaba, P.P. Irani, j. Leobe, Visualizing Casual Semantics using Animations, *IEEE, transactions on Visualization and Computer Graphics*, Vol 13, No 6, 2007.

[46] S. G. Eick, A. F. Karr, Visual Scalability, *IEEE Transactions on Visualization and Computer Graphics* , 2000

[47] G. Ellis, A. Dix, A Taxonomy of Clutter Reduction for Information Visualization, *IEEE, transactions on Visualization and Computer Graphics*, Vol 13 No. 6, pp 1216-1222, 2007

[48] Gay Lussac, J.L., Mem. de la Soc. d'Arcuel, 1809, translation 19, pp. 804 -- 818.

[49] Pander, Sadler, Ward, Shea, Euclidean Geometry, Cambridege Press, 2003

[50] Ronald Christensen, Analysis of variance, design and regression : applied statistical methods, Chapman & Hall, 1996.

[51] S. Axelsson, Intrusion Detection Systems: A Survey and Taxonomy, Chalmers Uni-versity, March 2000.

[52] S. Kumar, Classifcation and Detection of Computer Intrusions, PhD Thesis, Purdue University, August 1995.

[53] T. Kohonen, Self Organizing maps, Germany, 1995.

[54] O. Depren, M. Topallar, E. Anarim and M. K. Ciliz, An intelligent intrusion detection system (IDS) for anomaly and misuse detection in computer networks, Expert Systems with Applications 2005.

[55] J. Neter, Applied linear statistical model, McGraw Hill, Boston, 1996

[56] K. Mike, Applied linear regression modesl,McGraw Hill, Boston, 2005

[57] http://www.jibble.org/piespy/

[58] Greenberg, M. Jay, Euclidean and Non-Euclidean Geometries. Development and History, W. H. Freeman Company, 1974.

[59] Munzner, T. Interactive Visualization of Large Graphs and Networks, PhD Thesis, 2000.

[60] http://www.stanford.edu/group/sonia/papers/index.html

[61] Carla M. Dal Sasso Freitas1, Paulo R. G. Luzzardi, Ricardo A. Cava, Marco A. A. Winckler, Marcelo S. Pimenta, Luciana P. Nedel, Evaluating Usability of Information Visualization Techniques