To Cluster or not to Cluster

—

A Meta-Analytic Approach

Katharina A. Zweig*

May 31, 2008

Abstract

In the last years, a large number of graph clustering algorithms has been proposed that try to detect dense parts or *clusters* in a given network, e.g., [10, 14, 16, 32, 35, 34]. The resulting clusters are interpreted as functional modules, i.e., as a group of objects that have the same function within the network. This interpretation is implicitly justified by the assumption that most real–world networks are *local*, i.e., that edges between objects signify that they are *near* or *similar* to each other. Despite its importance, the term *locality* remains an intuitively well perceived but somehow elusive concept. In this article we pose the question of what a *local* graph is and discuss various findings in real–world networks. We extract a model of locality that can be easily tested if the set of objects is embedded in a metric space and we discuss what kind of locality is necessary in a real–world network in order to apply a clustering algorithm to it. The main result of this article is that the application of a clustering algorithm requires that a network is *strongly local*, a property that cannot be asserted without looking at the context in which a network is located. The article argues why, e.g., the Internet data on the level of autonomous systems and word–adjacency graphs are not likely to give reasonable results when clustered, and summarizes four characteristics of real-world networks that are required to be 'clusterable'.

1 Introduction

Almost 10 years ago, Watts and Strogatz built the basis for a new kind of science by introducing a new model for real–world networks, the so–called *small–world model* [39]. They showed that a graph family based on a so–called *local graph* and some random edges could account for some properties of real–world networks

small–world network models

*The author is supported by a scholarship from the Deutsche Akademie der Naturwissenschaften Leopoldina, with support from the Bundesministerium für Bildung und Forschung (BMBF, Germany) BMBF-LPD 9901/8-182.

that could not be explained by a purely random graph model. Following their article, a large number of different small–world models was presented, all of them based on a *local* and a random graph component [3, 7, 11, 20, 21, 23, 25]. The local graph component is, e.g., a ring in which every vertex is connected to its k-next neighbors [39], or any d-dimensional grid graph [20, 21, 25].

What is a local graph?

But what is a local graph in general? Intuitively speaking, a local graph is a graph in which objects are more likely to be connected if they are near to each other in a geometrical sense or if they are somewhat similar in a broader sense. In other words, in a local graph the objects represented by it can be characterized in various ways and this characterization can be used to define similarity between the objects. As an example Palla et al. point out that humans can be characterized in many different ways, and every characteristic defines a different set of people with which they come into contact [32]: the job will cause relationships with people working at the same place, as the membership in a sports-club will lead to relationships with other sportsmen. Most of the relationships will be to persons that are also near in a geographical sense, but some will be caused by a very special interest that connects experts from all over the world with each other. In the latter case, the *locality* cannot be interpreted by geographical distance but rather as a low distance in character space.

'Object & character space' vs. 'vertices & edges'

Thus, one the one hand we have the objects in character space that are similar to each other (or not). Then there is a *network generating process* that causes relations between the objects. These relations are then represented by vertices and edges in a network. Clustering algorithms have been developed to discover groups of densely connected vertices because it is assumed that they represent objects that are near in character space. In a way, the *character space* of a set of objects thus constitutes the **Platonic idea** of the real relationship between the objects. But as in Plato's allegory of the cave, the idea, i.e., the real character space in which objects are embedded, is often elusive for humans. Network data is now supposed to be something like the **Platonic shadow** of this *ideal characteristic space* (Fig. 1): if two objects in character space are next to each other, we assume that it is likely that they are connected in the network. Since the network is only a shadow we expect that some edges between near objects will not appear and that edges exist between objects that are not really near in the character space. This part of the networks is assumed to be the *random graph* part in the small–world models cited above.

Plato's allegory of the cave

clustering algorithms: basic assumption

More importantly for the idea of clustering, this assumption that near objects are connected by an edge is often considered to be a bi-implicaton (s. Fig. 2): it is assumed that if two objects are connected in a network then they are also near to each other in character space. Note that this is by no means obvious: A network could contain all the edges between near objects but add so many long-distance edges that it is not valid to assume that every edge signifies low distance. However, based on this bi-implication, a huge number of different clustering algorithms has been proposed in recent years [10, 14, 16, 32, 35, 34]. By taking only the structural information of the network (sometimes weighted, sometimes unweighted), these algorithms try to detect densely connected components of the graph that are only loosely connected with each other. It is

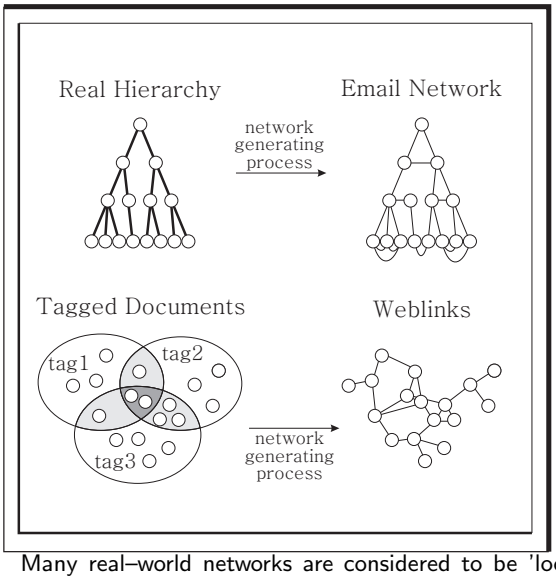

Figure 1: Many real–world networks are considered to be 'local', i.e., that the objects represented by the network are somewhat similar to each other (to different degrees), and the more similar they are to each other, the more likely it is that they are connected by an edge. The similarity between the objects could, e.g., be given by a hierarchy or by a tagging system (left hand side). It is expected that different kinds of relationships between the objects, e.g., the contacts by email or the weblinks between the objects, are dependent on the similarity of the objects in the space on the left hand side. I.e., we expect that there is a *network generating process*, e.g., the way people make contact, that favors the realization of those edges that are between near objects. The network is then considered to be a shadow of the character space, which can be rediscovered by clustering it (s. Fig. 2).

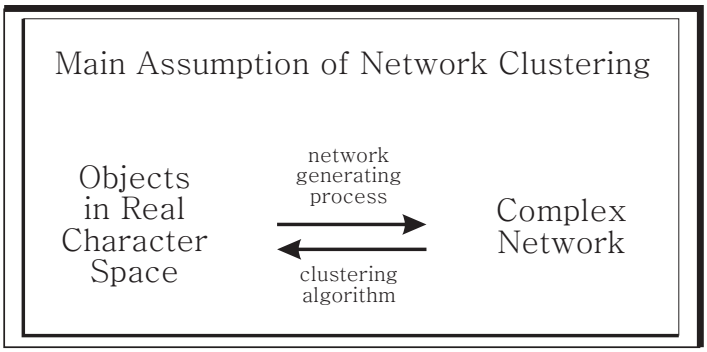

Figure 2: The main assumption for the application of clustering algorithms is that there exists a real character space in which the objects of interest are embedded and that the given network resembles this space by preferably linking objects that are near to each other in that space. It is then assumed that this *locality of edges* together with an uneven distribution of the objects in space generates a clustering structure in which dense subgraphs (clusters) emerge that are only loosely connected. These clusters can be detected by the various types of algorithms and are interpreted as functional modules of objects in the given system.

important to notice that these algorithms rely heavily on the assumption that an edge between objects signifies similarity between them in some meaningful character space. Thus, a network should only be clustered if it can be considered to be local.

How can we ensure locality?

To ensure locality of the graph it would thus be required to know the character space and the embedding of the objects in it beforehand. With this information it can then be checked whether most edges are between near objects. But of course, most often we use a clustering algorithm to find out about the character space! For example, clustering algorithms have been used successfully to find functional modules in protein–protein interaction networks [32], in metabolic networks [33], social groups of scientists, sports clubs, or friendship networks [29, 28, 27, 26], and groups of similar books in large warehouse data [9], to name but a few. And if one had had this information beforehand, a clustering would not have been necessary anymore.

Thus, the following questions remains: what if we do not have any contextual information, i.e., if we do not know the character space or if there is no reasonable distance measure beside the one given by the network? Are there special requirements of a graph to be meaningfully clustered? Are there examples of real–world networks that should not be clustered because they lack these requirements?

In the following we will first give some necessary definitions in Section 2 and then formalize the above given framework of locality in graphs in Section 3. We will then use this framework to analyze necessary conditions for applying a clustering algorithm to a given network in order to get meaningful results.

We will also present some examples of networks that are not likely to give good results if clustered. The article is summarized by open questions in Section 4.

2 Definitions

Let $G = (E, V)$ be a graph with $|V| = n$ vertices and $|E| = m$ edges. Note that $E \subseteq V \times V$ defines a *relation* between the vertices. Let $deg(v)$ denote the *degree* of v, i.e., the number of edges that v is contained in. Every vertex w with which v is connected by an edge $e = (v, w)$ is called a *neighbor* of v. Let $P_G(x, y)$ denote a *path in* G, i.e., a set of edges $\{e_1 = (x, x_1), e_2 = (x_1, x_2), \ldots, e_k(x_{k-1}, y)\}$. The *length* of a path is defined as the number of edges in it. The *distance* $d_G(x, y)$ between two vertices in G is defined as the minimal length of any path between them. If there is no such path, it is ∞ by definition. A graph $G' = (E', V')$ is called a *subgraph of* G if $V' \subseteq V$ and $E' \subseteq E$. A graph is called *connected* if there is a path between any two vertices. A subgraph G' of G is called *spanning* if it is a connected graph on all vertices of V (but does not necessarily contain all its edges).

In the $G(n, p)$ model, every of the $n(n-1)/2$ possible edges between n vertices is existent with probability p. In any given graph G with n vertices and m edges, it is easy to compute the so–called *edge density* $\rho(G)$ of G, defined as $2m/(n(n-1))$. A random graph $G(n, p)$ with $p = \rho(H)$ will then show approximately the same number of edges as H.

3 Locality in graphs

Our perception of the world is that our relationships to other persons or things are the more probable the nearer they are to us. This distance can be either interpreted in a purely geographical sense or in a more abstract way in a space defined by characteristics. Certainly, a network that is generated in a way where near things are more likely to be connected than distant ones will be perceived as a *local* graph. We will formalize this intuitive understanding:

Definition 3.1 (A first model of Locality) *Let O be a set of objects for which a distance measure $d : O \times O \to \mathbb{R}^+$ is defined. A graph on O is defined to be local, if $P(e = (x, y))$ is inversely proportional to some monotonically increasing function of $d(e = (x, y))$. $P(e)$ is called a local network generating process on O.*

Example 3.1 *With this definition we exclude random graphs from the set of local graphs since here $P(e) = p$ for all possible edges, independent of $d(e)$.*

As a shorthand we will use the term $d(e)$ to denote the distance between e's endpoints. Thus, given an embedding of some objects and a network between them, it is easy to check whether local edges are preferred over global edges. It has been shown for some real-world networks that indeed local edges are preferred:

local real–world networks

1. Gastner and Newman could show that in the design of commuter transport networks and sewage systems an intricate balance between the total geometric edge length and the travel time from any vertex to a center vertex along the network paths is achieved [15]. Moreover, in the networks they analyzed the total geometric edge length came close to the total edge length in the *minimal spanning tree* of the network, implying that every vertex prefers to be attached by its shortest edge to the growing network.

2. A second example is given by Frenken and van Oort who reviewed literature on the 'geometry of innovation' [13]. They summarize the findings described in the literature, and state that knowledge production and innovation are mainly achieved by groups whose members live in the same region. Additionally, they conducted a co-authorship analysis with respect to the affiliations of the authors. Two co-authors were considered to have a 'regional' cooperation if their affiliation lies in the same state. With this technique, they analyzed publications in two quite different scientific fields, namely 'aerospace engineering' and 'biotechnology and applied microbiology'. Their result is that scientific cooperations tend to be regional, although the trend has decreased in the last years due to cheaper communication, and that collaborations between academic and non-academic groups are more often regional than pure academic research.

3. Another interesting finding has been achieved by Yook, Jeong, and Barabási on the locality of the Internet, described on the level of routers and autonomous systems (AS) [40]. The authors used data collected by Govindan and Tangmunarunkit that mapped AS addresses to physical locations [17], and measured the probability $P(e)$ that edge e exists as a function of e's geometrical length $d(e)$. Their results clearly show that $P(e)$ is proportional to $1/d$. They explain this result with the costs of installing a physical link between routers that is assumed to be mainly growing linearly with its length.

4. Analyzing the email–contact network of a large company, Adamic and Adar found out that the probability that two persons had email contact (i.e., at least 6 emails) was proportional to $e^{-0.92h}$ where h denotes their distance in the hierarchy [1].

For many other real-world networks that exist between vertices with a fixed position it can be assumed with high certainty that most edges are local if the cost of building an edge is proportional to their distance. This argument is certainly valid for wires, tracks, streets, and also social relationships, although to a lesser extent as anyone can verify from his or her own acquaintanceship network.

Although preference of local edges seems to be settled for the above given networks, there are still classes of interesting networks out there where no distance function between the objects is readily available. We want to illustrate this important point with some examples:

presumably local real–world networks

1. A protein-protein interaction network represents the proteins of an organism by vertices, and two vertices are connected by an edge if the represented proteins interact biologically with each other. Since proteins often exert their function in the cell in tightly packed conglomerates, the understanding of protein–protein interaction networks is an important step to understand the time and space dependent functionality of cells. We will now explain why they are considered to be local graphs. It is assumed that these networks have at least partly evolved by duplicating certain parts of the genetic code that encodes proteins. Normally, the genetic code of a functional and vital protein is not allowed to change by much without losing its functionality. But if the code is duplicated, the genetic code of one copy can be mutated without harming the functionality of the original. Thereby the mutated protein may possibly lose some structural properties and gain others [30]. It might also interact with the original itself. This mechanism has been transformed into a dynamic network model that results in networks that are quite similar to the real ones, a good indication that the model captures the essential network generating process [37, 38]. If this mechanism models the evolution of protein-protein interaction correctly, then it is clear that at least some nearby proteins in the protein-protein interaction network are also similar to each other, e.g., on the level of their amino–acid sequence or their $3D$ structure. Under this model, we can assume a certain degree of locality in these networks. Still, the similarity of proteins is not unambiguously defined; measures range from similarity of the structure, to similarity of the amino-acid sequence they are made of, to the similarity of their function in the cell. This makes it very difficult to position proteins in any metric space or to define a coherent metric distance function between any two proteins such that all of these different similarities are captured.

2. We have a similar problem in metabolic networks. Here, all the small molecules produced by the set of enzymes of an organism are represented by vertices, and two vertices are connected if an enzyme catalyzes the transformation of one molecule into the other [12, 18]. Because the one is made of the other, it is clear that they share at least some structural properties and thus, metabolites with a low distance in this network can also be assumed to be structurally similar. And on the other hand, if two metabolites are very dissimilar than it is unlikely that a small number of enzyme catalyzed steps will transform the one into the other. Nonetheless, here it is also highly difficult to denote a metric distance function that captures the similarity between all pairs of metabolites.

3. Krebs [22] and Clauset [8] discuss co-purchasing networks of books where books sold by Amazon are represented by vertices. On the Amazon sites, every item's page contains links under the title: 'customers who bought this book also bought'. In a co-purchasing network two vertices are con-

nected by a (directed) edge if these links point from the one book to the other. As Clauset has shown in his article, a clustering of these networks reveals subgraphs consisting of very similar books, implying that edges are more likely between similar books. But—as can be seen in the comparison between any two libraries—there is no such thing as a unique categorization of books, and we know of no coherent quantitative measure that has been proposed to judge the similarity of two books.

4. As a last example we want to note the web's link structure [6, 2], modeled by networks where websites are represented as vertices and two vertices are connected by a (directed) edge if the one links to the other. Many algorithms have been proposed to harness this network structure to find those pages that are related to each other and to a given query, and their success is, without a doubt, amazing [5, 19, 24, 31]. We can thus safely assume that in this network those pages that are similar by content are also near each other in the network and that those pages that are near each other in the network are similar[1]. Still, it seems to be impossible to give a precise, metric distance function that quantifies the semantic similarity between any two websites.

In summary, there are many interesting real–world networks without a (known) embedding of the objects in character space but where the evolution of the network supports the assumption that most edges are between near or similar objects.

We will now analyze a couple of approaches to measure *localness* in graphs that rely only on structural information, i.e., on the adjacency matrix of a network. By abstracting the common properties of the measures proposed so far we will check whether the above given Definition 3.1 is sufficient to describe local graphs or needs refinement. The first approach to measure *localness* or the so–called 'cliquishness of a typical neighborhood (a local property)'[39] was the *clustering coefficient* $cc(v)$ of a vertex:

$$cc(v) = \frac{2e(v)}{deg(v) * (deg(v) - 1)}, \tag{1}$$

where $e(v)$ denotes the number of edges between neighbors of v. In other words, the clustering coefficient gives the probability that any two neighbors of v are connected by an edge themselves. In a random graph $G(n, p)$ this clustering coefficient is expectedly p (their edge density), but in real–world networks it is often magnitudes higher than expected by their edge density. It is intuitive that, if edges are more likely to occur between near objects that then any two neighbors of a vertex are also near to each other - and maybe near enough to make an edge between them likely. But still, there are many networks that are perceived as 'local graphs' that do not show a high clustering coefficient, e.g., grids or bipartite graphs. Thus, the idea of a clustering coefficient to measure

A local graph is expected to show transitivity

[1] Although there might exist different groups of websites concerning the same topic, e.g., if they are in different languages.

locality has been generalized in many ways, e.g., to counting the number of four-cycles in the neighborhood of a graph and comparing it with possible number of these structures [36]. Another approach is to define *locality* in the terms of *alternative paths*, i.e., an edge is considered to be local if there are at least k edge disjoint alternative paths of length at most l, as proposed by [3, 4]. Unfortunately, this measure is NP-hard to compute (i.e., not feasible) for any $k > 4$. Nonetheless, these measures have in common that they expect a local graph to be one that is transitive in the following way: If a, b are connected, we assume they are near to each other. If now b, c is also connected, we expect b is also near to c and that thus a cannot be too far away from c also (which might be near enough to make another edge probable).

Coming back to Definition 3.1, it can be noticed that this kind of transitivity is not explicitly required in the distance measure itself.

Example 3.2 *Consider the following example: let $P(e)$ be $d(e)^{-1}$, i.e., directly proportional to the inverse of the distance. This is certainly a local network generating process. Let now the distances between any two pairs of nodes be either 1 or ∞. It follows, that all edges between pairs of vertices with distance 1 are in the graph and no edges between pairs of vertices with distance ∞. If such an embedding or distance measure is where we find the vertices in, it might be that the resulting network does not show any clusters: Let $G(n, p)$ be a random graph and let it be embedded s.t. $d(v, w) = 1$ iff $(v, w) \in E$ and $d(v, w) = \infty$, else. Although now the random graph can be considered to be local with respect to the given embedding, it is certainly not meaningfully clusterable.*

This shows that we have to make requirements with respect to the embedding of the objects in character space. If the objects do not cluster in the full character space, i.e., if there are no discernible groups of objects that are more similar to each other than to other objects, also the local network generated out of this embedding will not be clusterable. We think that the following type of networks belongs to this group:

Example 3.3 *A (simple) word adjacency graph (waG) is based one or more pieces of literature where the words are represented by vertices and two vertices are connected if the corresponding words are next to each other in at least one sentence. Is it useful to apply a clustering algorithm to this network, i.e., will we find groups of densely connected words? We define $d(w_1, w_2)$, the distance between any two words w_1, w_2, to be $(\#adj(w_1, w_2))^{-1}$ where $\#adj(w_1, w_2)$ denotes the number of times they are standing next to each other in the given text (and $d(w_1, w_2) = \infty$ if $\#adj(w_1, w_2) = 0$). Then, we define $P(e) = d(e)^{-1}$. The problem with this embedding is that it is not likely to produce groups of words that are similar to each other with respect to this distance measure, i.e., if $d(w_1, w_2) = x \& d(w_2, w_3) = y$ it is likely that $d(w_1, w_3) > x + y$. The sentence "In the beginning..." makes two pairs of words (in/the, the/beginning), and certainly the pair (in/the) will be found very often but still it is unlikely to find (in/beginning) somewhere in the text.*

Thus, the network shadows a character space that is not likely to contain many groups of objects with a certain, pairwise similarity. The problem with the Definition 3.1 is that it allows any kind of distance measure. We will thus sharpen the definition by requiring the distance measure $d : O \times O \to \mathbb{R}^+$ to obey the triangle equality[2]. By furthermore requiring a symmetric distance measure, i.e., $d(x, y) = d(y, x)$, we are in other words requiring the following:

Definition 3.2 (Local Network Generating Process) *Let O be a set of objects embedded in a metric space S, i.e., there exists a distance measure $d : O \times O$ such that*

1. $d(x, y) \geq 0$;
2. $d(x, y) == 0$ *iff* $x == y$;
3. $d(x, y) = d(y, x)$;
4. $d(x, z) \leq d(x, y) + d(y, z)$.

Then $\mathcal{P}_S$ is a local network generating process in S if it produces a relation $E(P(e)) \subseteq O \times O$ such that the probability $P(e)$ that an edge e exists is inversely proportional to some monotonically increasing function $f(d(e))$ of the distance between its endpoints.

A graph $G = (O, E(P_S(e))$ is called a local graph.

It is a very important observation that we require certain conditions on both sides of a network generating process, in the object/character space and the network. We require that the similarity measure is at least in a way transitive and we require that the more similar two objects are, the higher the probability that they are connected by an edge in the network.

Although this sounds like a very reasonable model, it will need even stronger constraints to make the application of clustering algorithms possible:

Example 3.4 *Suppose that the objects O are distributed in a way such that the number of objects with distance d to v is described by $N_v(d)$. If then $\lim_{d\to\infty} N_v(d)/f(d)$ is larger than a constant, i.e., if N_v grows asymptotically as least as fast or faster than $f(d)$, there will be* **the same number or even absolutely more** *edges to distant objects than to near objects. An example for this is the Internet: as cited above, Yook et al. showed that a link between two routers in distance d is proportional to $1/d$. Since routers are distributed over a two-dimensional area it can be expected that $N_v(d)$ is roughly proportional to d. This implies that $f(d)$ and $N(d)$ are both linear in d, i.e., their ratio is some constant c. This approximation implies that v is connected to a constant number c of routers in* **every** *distance d. Thus, picking any edge e of the Internet graph, it has the same probability to connect two servers in distance d_1 or d_2 for all possible distances! This approximative calculation implies that a clustering*

[2]Note that the choice of the triangle equality is not the only possible way to ensure transitivity, but a natural one in many settings.

algorithm will not be able to find groups of servers that are in the same geographical area since an edge does **not** *imply geographical nearness between the servers.*

We will thus introduce the following two definitions to classify local graphs:

Definition 3.3 *1. A graph is called strongly local if – given any edge of the graph – it is more or at least as likely that the connected vertices are in distance d than in distance d' for all distances $d < d'$.*

2. A graph is called weakly local if two vertices in distance d are more likely to be connected than two vertices in distance d' for all distances $d < d'$.

Although both definitions sound very similar, the important difference between them is that in a strongly local graph absolutely more edges are local than distant. In a weakly local graph the probability of all short edges to exist is higher than that of long edges - but if there are many more long edges than short ones there could be absolutely more long edges than short ones. Now, it becomes clear that we do not only require locality for a graph to show a meaningful clustering. In order to apply a clustering algorithm to a given real–world network, we should have reasonable hopes that the given graph is *strongly local*, i.e., that most of the edges in the graph are between near or similar vertices. In a weakly local graph that is not strongly local most of the edges will be between distanct vertices just because there are more possible edges with a long distance between the objects.

A last problem can come up when the network is strongly local with respect to some character space but its clustering results are interpreted with respect to another one:

Example 3.5 *Proteins can be in various relationships, e.g., they can interact with each other, they can have a similar gene sequence, they can be produced under the same circumstance, etc. Each of these relations gives rise to a network that can be clustered. One of the most important questions in protein chemistry is that of annotation, i.e., what function the protein has. Thus, the character space that should be explored is that of functionality, and two proteins can be called similar if they have a similar function in the cell. The question is now which of the relations cited above shadows this character space best. At the moment, results of protein-protein-interaction networks seem to be quite promising [32] but the predictions will have to be tested in the laboratory.*

4 Discussion

In this article we have discussed a general model of when clustering algorithms can be applied to networks. We have introduced a Platonic model that describes the relationship between the ideal character space in which objects are embedded and a network between them. We have argued, that networks are only (meaningfully) clusterable if the embedding in space allows for groups of

objects that are pairwise more similar to each other than with other objects. We have further argued that the network generating process must be strongly local such that most edges in the resulting networks are between objects that are similar in the character space that is to be explored.

In summary, before a clustering algorithms is applied to a real–world network it should be checked that the following four conditions are satisfied:

1. The network generating process that transforms the character space into a network is local (Example 3.2).
2. The embedding of the objects in the underlying character space is likely to give rise to clusters, i.e., it can be expected that there are groups of objects that are pairwise more similar to each other than to other objects (Example 3.3.
3. Moreover, it can be expected that the network generating process is strongly local, i.e., an edge in the network is likely to signify similarity between the objects (Example 3.4).
4. The network shadows the intended character space (Example 3.5.

References

[1] Lada Adamic and Eytan Adar. How to search a social network. *Social Networks*, 27(3):187–203, 2005.

[2] Albert-László, Hawoong Jeong, and Réka Albert. The diameter of the world wide web. *Nature*, 401:130, 1999.

[3] Reid Andersen, Fan Chung, and Lincoln Lu. Analyzing the small world phenomenon using a hybrid model with local network flow. In *Proceedings of the WAW 2004*, LNCS 3243, 2004.

[4] Reid Andersen, Fan Chung, and Linyuan Lu. Drawing power law graphs. In *Proceedings of the 12th Symposium on Graph Drawing (GD'04)*, 2004.

[5] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, 1998.

[6] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stat, Andrew Tomkins, and Janet Wiener. Graph structure in the web. *Computer Networks*, 33:309–320, 2000.

[7] Fan Chung and Linyuan Lu. The small world phenomenon in hybrid power law graphs. In *Complex Networks (E. Ben-Naim, H. Frauenfelder, Z. Toroczkai (eds.))*, pages 91–106, 2004.

[8] Aaron Clauset. Finding local community structure in networks. *Physical Review E*, 72:026132, 2005.

[9] Aaron Clauset, Mark E.J. Newman, and Christopher Moore. Finding community structure in very large networks. *Physical Review E*, 70:066111, 2004.

[10] Imre Derényi, Gergely Palla, and Tamás Vicsek. Clique percolation in random networks. *Phys. Rev. Lett.*, 94:160202, 2005.

[11] Sergei N. Dorogovtsev and Jose F.F. Mendes. Exactly solvable analogy of small-world networks. *Europhys. Lett.*, 50:1–7, 2000.

[12] David A. Fell and Andreas Wagner. The small world of metabolism. *Nature Biotechnology*, 18:1121–1122, 2000.

[13] Koen Frenken and Frank G. van Oort. The geography of research collaboration in US aerospace engineering and US biotechnology & applied microbiology. In *Conference of the Regional Studies Association*, 2003.

[14] Marco Gaertler. *Network Analysis: Methodological Foundations*, chapter Clustering, pages 178–215. Springer-Verlag, 2005.

[15] Michael T. Gastner and Mark E.J. Newman. Shape and efficiency in spatial distribution networks. *Journal of Statistical Mechanics: Theory and Experiment*, page P01015, September 2004.

[16] Michelle Girvan and Mark E.J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99:7821–7826, 2002.

[17] R. Govindan and H. Tangmunarunkit. In *Proceedings of the IEEE INFOCOM'00, Tel Aviv*, pages 1371–1380, 2000.

[18] H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, and A.-L. Barabsi. The large-scale organization of metabolic networks. *Nature*, 400:107, 2000.

[19] Jon Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[20] Jon Kleinberg. Navigation in a small world. *Nature*, 406:845, 2000.

[21] Jon Kleinberg. The small-world phenomenon: An algorithmic perspective. In *Proceedings of the 32nd ACM Symposium on Theory of Computing*, 2000.

[22] Valdis Krebs. The social life of books. http://www.orgnet.com/booknet.html.

[23] Katharina A. Lehmann, Hendrik D. Post, and Michael Kaufmann. Hybrid graphs as a framework for the small-world effect. *Physical Review E*, 73:056108, 2006.

[24] Ronny Lempel and Shlomo Moran. The stochastic approach for link-structure analysis (SALSA) and the TKC effect. *Computer Networks: The International Journal of COmputer and Telecommunication Networking*, 33:387–401, 2000.

[25] M. E. J. Newman and D. J. Watts. Scaling and percolation in the small-world network model. *Phys. Rev. E*, 60:7332–7342, 1999.

[26] Mark E.J. Newman. Who is the best connected scientist? a study of scientific coauthorship networks. arXiv: cond-mat/0011144, November 2000.

[27] Mark E.J. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences, USA*, 98(2):404–409, 2001.

[28] Mark E.J. Newman. Fast algorithm for detecting community structure in networks. *Phys. Rev. E*, 69:066133, 2004.

[29] Mark E.J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the USA*, 103:8577–8582, 2006.

[30] S. Ohno. *Evolution by Gene Duplication*. Springer Verlag, Berlin, 1970.

[31] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Manuscript, 1999.

[32] G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435:814, 2005.

[33] E. Ravasz, A.L. Somera, D.A. Mongru, Z.N. Oltvai, and A.-L. Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 297:1551–1553, 2002.

[34] Jörg Reichardt and Stefan Bornholdt. Partitioning and modularity of graphs with arbitrary degree distribution. arXiv:cond-mat/0606295, Juni 2006.

[35] Jörg Reichardt and Stefan Bornholdt. When are networks truly modular? arXiv:cond-mat/0606220, Juni 2006.

[36] Garry Robins, Philippa Pattison, and Jodie Woolcock. Small and other worlds: Global network structures from local processes.

[37] Ricard V. Solé, Romualdo Pastor-Satorras, Eric Smith, and Thomas B. Kepler. A model of large-scale proteome evolution. *Advances in Complex Systems*, 5:43–54, 2002.

[38] Alexei Vázquez, Alessandro Flammini, Amos Maritan, and Alessandro Vespignani. Modeling of protein interaction networks. *ComPlexUs*, 1:38–44, 2002.

[39] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, June 1998.

[40] Soon-Hyung Yook, Hawoong Jeong, and Albert-László Barabási. Modeling the internet's large-scale topology. *Proceedings of the National Academy of the Sciences, USA*, 99(21):13382–3386, 2002.