# A Quality Measure for Multi-Level Community Structure

Maylis Delest, Jean-Marc Fédou, Guy Melançon

# A Quality Measure for Multi-Level Community Structure

Maylis Delest
LaBRI UMR CNRS 5800
Bordeaux, France
Maylis.Delest@labri.fr

Jean-Marc Fédou
I3S UMR CNRS 6070
Nice Sophia Antipolis, France
Jean-Marc.Fedou@unice.fr

Guy Melançon
LIRMM UMR CNRS 5506
Montpellier, France
Guy.Melancon@lirmm.fr

## Abstract

*Mining relational data often boils down to computing clusters, that is finding sub-communities of data elements forming cohesive sub-units, while being well separated from one another. The clusters themselves are sometimes terms "communities" and the way clusters relate to one another is often referred to as a "community structure". We study a modularity criterion $MQ$ introduced by Mancoridis* et al. *in order to infer community structure on relational data.*

*We prove a fundamental and useful property of the modularity measure $MQ$, showing that it can be approximated by a gaussian distribution, making it a prevalent choice over less focused optimization criterion for graph clustering. This makes it possible to compare two different clusterings of a same graph as well as asserting the overall quality of a given clustering relying on the fact that $MQ$ is gaussian. Moreover, we introduce a generalization extending $MQ$ to hierarchical clusterings of graphs which reduces to the original $MQ$ when the hierarchy becomes flat.*

## 1 Introduction

Network science has now become a major research area involving scientists from various disciplines: sociologists, computer scientists, physicists and mathematicians now meet regularly to exchange ideas on issues related to the study of networks [3, 5]. Questions concern graph models and their properties, algorithms and related implementation issues, empirical study trying to validate models, etc.

Networks indeed appear as natural models in numerous application domains. People participating to a same social activity, companies competing or collaborating in a given industrial sector, routers exchanging packets over the internet are all examples of networks that can be modeled using graphs. They form a network because of the interactions taking place between the different actors: people, companies or routers. Networks are commonly used in biology to model protein interaction when addressing the problem of finding functional relationships between biological objects. Co-occurrence of terms or concepts in text or hypermedia documents provide a fruitful strategy to explore large information space. More generally, networks also appear as a useful tool to explore data in context where relations must be induced by interpreting the available data. Computer science contributes to this vivid research field by providing algorithms capable of searching a large network hoping to identify "natural" clusters or communities describing its overall structure. Once a sub-community has been identified, the analyst will typically pursue a more detailed inspection of its own dynamics.

Being able to find the intrinsic community structure of a relational dataset is of interest to data miners. Indeed, once communities have been identified, the original set can be reduced to a quotient making explicit the relations between them, thus enabling the analyst to identify higher-level pattern in the data. The ability of assessing of the intrinsic quality of a community structure is an issue we wish to address in this short note. More precisely, we promote the use of an index quantifying the quality of a graph clustering introduced by *Mancoridis et al.* [10]. We show that *Mancoridis et al.*'s quality index possess important statistical properties making it a more relevant choice over other possibilities.

## 2 Identifying network communities: quality measure of community structures

### 2.1 Mancoridis *et al.*'s $MQ$

The problem of finding communities or "natural clusters" in a graph had been adressed by Mancoridis *et al.* [10] in the context of software reverse-engineering where communities correspond to logical units of programs. Their approach made use of a map computing the "modularity quality" ($MQ$) of a clustering in terms of internal cohesion and outer communications between units. Their method mainly consisted in seeing $MQ$ as an optimization criterion. They used classical approaches such as genetic algorithms or hill climbing in order to find community structures with maximum modularity quality $MQ$. Auber *et al.* [2] later used this $MQ$ criterion in order to find communities in small world networks (social networks). They used $MQ$ to tune a threshold value filtering edges of the graph (thus maximizing $MQ$ along a one-dimensional parameter), leading to a fragmentation of the graph into connected components from which they induced a clustering. Recursively using $MQ$ on sub-communities, Auber *et al.* obtained a multi-level decompositions of graphs. They also showed how this hierarchical decomposition can be used as a visual metaphor for exploring large graphs.

In order to define Mancoridis' $MQ$, we need to introduce some notations. Let $G = (V, E)$ with $n = |V|$ be a simple graph[1] over a set $V = \{v_1, \ldots, v_n\}$. Denote as usual by $N_G(v)$ the neighborhood of $v$ in $G$, that is the set of nodes connected to $v$ by an edge in $E$.

Let $\mathbf{C}$ be a *clustering* $(C_1, \ldots, C_k)$ where the subsets $C_i \subset V$ are pairwise disjoint and sum up to $C_1 \cup \cdots \cup C_k = V$. A clustering is also sometimes called a *set partition* of the set $V$. We shall need notations describing the size of various neighborhoods with respect to $\mathbf{C}$. To this end, we introduce two matrices $\mathbf{C} = (c_{i,p})$, $\mathbf{D} = (d_{i,q})$ with:

$$c_{i,p} = \begin{cases} 1 & \text{if } v_p \in C_i \\ 0 & \text{otherwise} \end{cases} \qquad d_{i,q} = |N_G(v_q) \cap C_i|$$

Note that we abuse notations and write $\mathbf{C}$ for both the clustering and the matrix encoding the membership relation of nodes $v_p$ to clusters $C_i$. Entries of row $i$ of the matrix $\mathbf{C}$ correspond to nodes of the subset $C_i$. By definition, each column of $\mathbf{C}$ contains a single entry equal to 1, all others being equal to 0, and we have $\sum_p c_{i,p} = |C_i|$. As for matrix $\mathbf{D}$, the entry $d_{i,q}$ equals the number of neighbors of node $v_q$ belonging to cluster $C_i$, so we have $\sum_i d_{i,q} = d_G(v_q)$ (the degree of node $v_q$ in $G$). Each row of matrix $\mathbf{C}$ or $\mathbf{D}$ can be thought of as an $n$-dimensional vector. Denote by $\langle \cdot\,, \cdot \rangle$ the (symmetric) bilinear form computing the usual scalar product:

$$\langle X, Y \rangle = \sum_{p=1}^{N} x_p y_p.$$

**Lemma 2.1** *The number of edges connecting nodes between $C_i$ and $C_j$ ($i \neq j$) can be computed as: $(\langle C_i, D_j \rangle + \langle C_j, D_i \rangle)/2$.*

*When $i = j$, the number of edges connecting nodes in $C_i$ is given by $\langle C_i, D_i \rangle / 2$.*

Note however, that in practice the number of edges between $C_i$ and $C_j$ can be computed as $\langle C_i, D_j \rangle = \sum_{p=1}^{n} c_{i,p} d_{j,p}$ since we have $\langle C_i, D_j \rangle = \langle C_j, D_i \rangle$.

**Definition 2.1** *(See [10]) Define*

$$\begin{aligned} MQ^+(G; \mathbf{C}) &= \frac{1}{k} \sum_{i=1}^{k} \frac{\langle C_i, D_i \rangle / 2}{\binom{|C_i|}{2}} \quad , \\ MQ^-(G; \mathbf{C}) &= \frac{1}{\binom{k}{2}} \sum_{i<j} \frac{\langle C_i, D_j \rangle}{|C_i||C_j|} \quad . \end{aligned}$$

*and set $MQ(G; \mathbf{C}) = MQ^+(G; \mathbf{C}) - MQ^-(G; \mathbf{C})$.*

We refer to the term $MQ^+(G; \mathbf{C})$ as a positive contribution to $MQ$ and to the term $MQ^-(G; \mathbf{C})$ as a negative contribution. Indeed, a "good" clustering should favor edges internal to clusters and should try to minimize the number of edges connecting nodes of different clusters.

The ratio computed by the term $\frac{\langle C_i, D_i \rangle / 2}{\binom{|C_i|}{2}}$ measures how close cluster $C_i$ is to a clique – it can actually be seen as an extension of Watt's clustering index [12, 13] to clusters of the graph $G$. Conversely, the term $\frac{\langle C_i, D_j \rangle}{|C_i||C_j|}$ indicates how close edges between $C_i$ and $C_j$ are to a complete bipartite graph. This ratio could be interpreted as a dissimilarity between the sets $C_i$ and $C_j$, by analogy to the *link index* introduced by Guha *et al.* [8] as part of the ROCK clustering algorithm.

*Example*. Consider the graph shown in Figure 1 (borrowed from [7]). Nodes represent members of a karate club and edges models acquaintances between them. Nodes have been divided into two different clusters and are marked either as ovals or squares. Computing $MQ$ for this graph and this clustering gives a quality index of $0.1742$ which is actually quite good since approximately only 3% of all partitions have an $MQ$ value above $0.174$ (assuming an average value of $-0.2$ and standard deviation of $0.2$, see [4]).

[1] That is, $G$ is undirected and contains no self-loop.

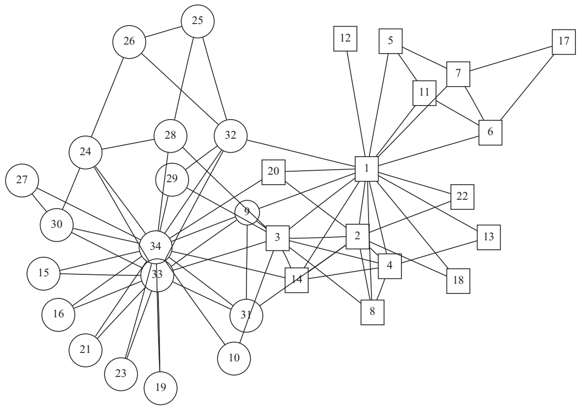

**Figure 1. Zachary's karate club graph borrowed from [7]. Nodes belong to either of two clusters according to whether they are shown as circles or squares.**

## 3 Gaussian approximation for $MQ$

The quality index $MQ$ has several advantages over other possible cost functions. First observe that $MQ$ varies over the finite interval $[-1, 1]$. Indeed, $MQ(G; \mathbf{C}) = 1$ when all clusters $C_i$ are cliques (graphs containing all possible edges) and that no edges connect nodes of different clusters. Similarly, $MQ(G; \mathbf{C}) = -1$ when clusters contain no internal edges while pairs of clusters $C_i$ and $C_j$ form complete bipartite graphs. This already makes it easier to compare different clusterings of a same graph. Other min-cut cost function often simply "count" the number of edges (or their weights) linking distinct clusters, thus varying over an unbounded domain (see the surveys [1, 9]). Also, $MQ$ takes the size (number of nodes) of clusters into account thus requiring a cluster to reach a reasonably high "density", as opposed to other cost function which only require a cluster to have numerous edges without respect to its number of nodes (this is the case for the quality index $Q$ introduced by Girvan and Newman [11], for instance).

We now show that $MQ$ can be approximated by a gaussian distribution providing a criterion to compare clusterings of graphs on a common scale, but most of all to objectively compare clustering heuristics. The proof relies on two variations of the central limit theorem we now state.

**Theorem 3.1** *Let $(X_i)_{i\geq 1}$ be a sequence of* independent *random variables. Then the random variable*

$$\frac{X_1 + \cdots + X_k}{k}$$

*converges towards a gaussian random variable $N(\mu, \sigma)$ as $k \rightarrow \infty$.*

The following "variation" however does not require that the variables be independent. It is often referred to as the *de Moivre-Laplace* theorem:

**Theorem 3.2** *Let $(X_i)_{i\geq 1}$ be a sequence of random variables obeying the same probability distribution, all having the same mean value $\bar{X}_i = \mu$ and standard deviation $\sigma$. Then the random variable*

$$\frac{1}{\sqrt{k}}(\frac{X_1 + \cdots + X_k - k\mu}{\sigma})$$

*converges towards a centered and normalized gaussian random variable $N(0, 1)$ as $k \rightarrow \infty$. As a consequence, the variable*

$$\frac{X_1 + \cdots + X_k}{k}$$

*can be approximated by a gaussian distribution with mean $\mu$ and standard deviation $\sigma/\sqrt{k}$.*

(For more details on these classical results, the reader should consult basic textbooks such as [6]).

**Theorem 3.3** *Consider $MQ$ as a random variable depending on both a graph $G = (V, E)$ with $n = |V|$ and a clustering $\mathbf{C} = (C_1, \ldots, C_k)$. Then $MQ$ can be approximated by a gaussian distribution as $n \rightarrow \infty$.*

We shall sketch proofs showing that both $M^+(G; \mathbf{C})$ and $M^-(G; \mathbf{C})$ can be approximated by a gaussian distribution. Now assume these terms can indeed be seen as random variables. They furthermore are independent since they rely on disjoint subsets of edges of the underlying graph $G$. Hence, the theorem will follow since the sum of any two independent gaussian random variables is again a gaussian random variable.

Now, each term of Definition 2.1 can be considered as a random variable taking its value from a graph $G$ and a cluster $C_i$. We shall see that each of these terms can be approximated by a gaussian distribution (see the following lemmas and corollaries). Again, these random variables are independent because they rely on disjoint subsets of edges in $G$, thus their sum provides a gaussian approximation of $M^+(G; \mathbf{C})$, by virtue of Theorem 3.1. The same argument can be repeated with Eq. (1) for $M^-(G; \mathbf{C})$.

**Lemma 3.1** *Let $G = (V, E)$ be a graph with $V = \{1, \ldots, n\}$ so that edges correspond to pairs $\{p, q\}$ of distinct integers ($p, q \in V$ and $p \neq q$). For each of these pairs $\{p, q\}$ define the random variable $X_{p,q}$ as:*

$$X_{p,q}(G) \quad = \quad \begin{cases} 1 & if\ \{p, q\}\ is\ an\ edge\ in\ E \\ 0 & otherwise \end{cases}$$

*Then the variable $X = \frac{\sum_{p,q} X_{p,q}}{\binom{n}{2}}$ converges towards a gaussian distribution as $n \rightarrow \infty$.*

First note that the number of random variables $X_{p,q}$ associated with pairs of distinct integers $\{p, q\}$ is exactly $\binom{n}{2}$. Obviously, all variables $X_{p,q}$ obey the same distribution, have the same mean $\mu_{p,q} = 1/2$ and standard deviation $\sigma_{p,q} = 1/2$. Indeed, the edges of a graph $G$ can be put into one-to-one correspondence with subsets of all possible pairs of distinct integers. Observe that half of these subsets do contain a given pair $p, q$ while the other half does not, from which we deduce $\bar{X}_{p,q} = 1/2$ and $\sigma_{p,q} = 1/2$. By virtue of Theorem 3.2, the variable $X$ can be approximated by a gaussian distribution.

**Corollary 3.1** *The expression $\frac{E(C_i)}{\binom{|C_i|}{2}}$ can be considered as a random variable depending of a graph G (where $\mathbf{C}$ is assumed to be given) and as such can be approximated by a gaussian distribution.*

The variable indeed depends on the subgraph induced from $G$ on the subset $C_i$. Moreover, we have:

$$X(G|_{C_i}) = \sum_{\substack{p,q \in C_i \\ p \neq q}} X_{p,q}(G|_{C_i}) = \frac{E(C_i)}{\binom{|C_i|}{2}}$$

where the variables $X_{p,q}$ are defined as in Lemma 3.1. Consequently, lemma 3.1 applies and $X$ can be approximated by a gaussian random distribution.

The same type of argument can be repeated for $MQ^{-}(G;)$ relying on bipartite graphs $K_{r,s} = (V, E)$ and random variables

**Lemma 3.2** *Let $K_{r,s} = (V, E)$ denote the complete bipartite graph. That is, $V = V_1 \oplus V_2$ with $V_1 = \{1, \ldots, r\}$ and $V_2 = \{1, \ldots, s\}$ and edges in $E$ correspond to pairs $\{p, q\}$ with $p \in V_1$ and $q \in V_2$. For each of these pairs $\{p, q\}$ define the random variable $Y_{p,q}$ as $Y_{p,q}(G) =$*
$$\begin{cases} 1 & if\ \{p,q\}\ is\ an\ edge\ in\ E \\ 0 & otherwise \end{cases}$$

The proof mimics that of Lemma 3.1. Again, note that the number of random variables $X_{p,q}$ associated with pairs $\{p, q\}$ such that $p \in V_1$ and $q \in V_2$ is exactly $r \cdot s$. The same argument as that used in Lemma 3.1 shows that all variables $Y_{p,q}$ have the same mean $\mu_{p,q} = 1/2$ and standard deviation $\sigma_{p,q} = 1/2$. By virtue of Theorem 3.2, the variable $Y$ can be approximated by a gaussian distribution.

**Corollary 3.2** *The expression $\frac{E(C_i,C_j)}{|C_i||C_j|}$ can be considered as a random variable depending on a graph G (where $\mathbf{C}$ is assumed to be given) and as such can be approximated by a gaussian distribution.*

## 4 A multilevel quality measure for hierarchical clustering

We now embark on defining what we call a *hierarchical clustering of a graph*. The general idea is to have a set of nested clusters building into a hierarchy and covering the whole graph. Obviously, we aim at generalizing $MQ$ and define a quality measure taking the whole hierarchy of clusters into account.

Let us first define a rooted tree as a *connected* simple graph $T = (W, F)$ which moreover does not contain any cycle, with a distinguished node $r \in W$ called its *root*. An *ancestor* of a node $x$ is a node $z$ situated on the path connecting $x$ to the root. As a consequence, any node but $r$ has at least one ancestor. A node $y$ is a *descendant* of a node $x$ when $x$ is part of the pat connecting $y$ to the root. The *parent* node $p(x)$ of a node $x$ is its closest ancestor – the ancestor sitting at distance one from $x$. Any two nodes $x, y$ necessarily have a *common ancestor* (which can be either $x$ or $y$). The *child nodes* of a node $x$ are its closest descendants. A leaf is a node with no child. The set of leaf nodes of a tree $T$ will be denoted as $L(T)$. Finally, a subtree $T_x$ can be induced from any node $x \in W$. It is obtained by taking the subgraph induced from the set of all descendants of $x$, including $x$ itself acting as a root for $T_x$.

A *hierarchical clustering* of a graph $G = (V, E)$ is a tree $T = (W, F)$ where $W \subset 2^V$, that is the nodes of the tree are subsets of elements in $V$. We shall denote subsets of $V$ as $x$ or $y$. We impose that a node $x$ be equal to the set union of its child nodes $x = y_1 \cup \cdots \cup y_m$, and that they be distinct $y_1 \cap \cdots \cap y_m = \emptyset$. We also require that subsets attached to leaves of $T$ cover the set $V$. That is, leaves of the tree $T$ correspond to clusters of nodes in the graph $G$. Our definition differs from the more usual one where leaf nodes correspond to nodes $v \in V$ (or, more precisely, singletons $\{v\}$). Note also that any node $v \in V$ belongs to a *unique* leaf node $x \subset V$ in $T$.

By convenience, we will simply say that the couple $(G; T)$ is a hierarchically clustered graph. Let $x$ denote the root of the tree $T$ and $x_1, \ldots, x_k$ denote the child nodes of $x$ in $T$. Denote by $T_{x_1}, \ldots, T_{x_k}$ the subtrees rooted at $x_1, \ldots, x_k$. Observe that the subsets $C_1 = L(T_{x_1})$, $\ldots$, $C_q = L(T_{x_q})$ induced from the leaves pending at the subtrees induce a clustering of $G$. Moreover, the subtrees $T_{x_1}, \ldots, T_{x_k}$ together with the induced subgraph $G(C_1)$, $\ldots$, $G(C_k)$ correspond to hierarchically clustered graphs $(G(C_1); T_{x_1}), \ldots, (G(C_k); T_{x_k})$.

The quality measure we wish to define follows the idea that an “internal” edge (connecting nodes of a same cluster) should be encouraged to sit as deep as possible in the tree. Similarly, the penalty assigned to external edges (connecting nodes of different clusters) should be somewhat correlated to the tree distance between the two clusters. This

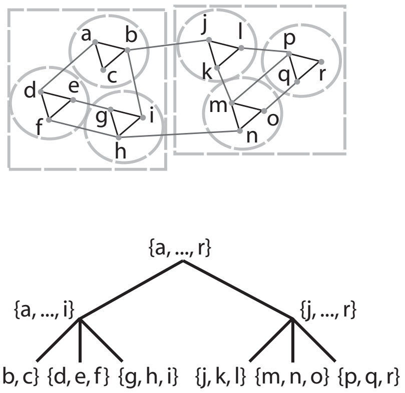

**Figure 2. A clustered graph $G = (V, E)$ is a graph equipped with a tree structure whose leaves are *distinct* subsets $x \subset V$ covering $V$. The cluster structure is indicated by the dashed grayed regions.**

is accomplished by assigning weights to edges, cumulating (positive or negative) values correlated with the depth at which the end nodes sit. This is accomplished by recursively defining a multilevel $mMQ$ index as follows:

**Definition 4.1** *Let $(G;T)$ be a hierarchially clustered graph and let $0 < q < 1$ be any real number. The multilevel modularity quality $MQ(q;G;T)$ is defined in terms of the hierarchically clustered subgraphs $(G(C_1);T_{x_1})$, ..., $(G(C_k);T_{x_k})$ as follows:*

$$MQ^+(q;G;T)$$

$$= \frac{1}{k}\sum_{i=1}^{k} \frac{\langle C_i, D_i\rangle/2}{\binom{|C_i|}{2}}\big(1 + q\cdot MQ^+(q;G(C_i);T_{x_i})\big),$$

$$MQ^-(q;G;T)$$

$$= \frac{1}{\binom{k}{2}}\sum_{i<j} \frac{\langle C_i, D_j\rangle}{|C_i||C_j|}\big(1 + q\cdot MQ^-(q;G(C_i);T_{x_i}\big)$$

$$\cdot\big(1 + q\cdot MQ^-(q;G(C_j;T_{x_j}\big).$$

*Finally,*

$$MQ(q;G;T) = MQ^+(q;G;T) - MQ^-(q;G;T).$$

As one can observe, for each edge of the graph, the recursion actually cumulates powers of $q$ depending on its depth with respect to the hierarchy tree $T$. The total contribution of an edge to $MQ(q;G;T)$ varies according to the depth at which it starts being external to clusters nested more deeply in the hierarchy.

**Proposition 4.1** *Let $(G,T)$ be a hierarchical clustering and denote by $\mathbf{C} = C_1, \ldots, C_k$ the clustering of the graph $G$ induced from the subgraphs $G(L(T_{x_1}))$, ..., $G(L(T_{x_k}))$ where $x_1, \ldots, x_k$ are the child nodes of the root $x$.*

*Then we have $MQ(G;\mathbf{C}) = MQ(0;G;T)$.*

Indeed, setting $q = 0$ is equivalent to flattening the hierarchy thus only seeing edges as acting between clusters of depth 1.

**Theorem 4.1** *Given a real number $0 < q < 1$, the multilevel quality measure $MQ(q;\bullet;\bullet)$ varies over the interval $(-\frac{1}{1-q}, \frac{1}{1-q})$. Moreover, it can be approximated by a gaussian distribution.*

The firt part of the statement of the theorem follows by observing that both the positive and negative contributions of an edge (its associated coefficients in $MQ^+(q;\bullet;\bullet)$ and $MQ^-(q;\bullet;\bullet)$ respectively) are bounded by the series $\sum_{i\geq 0} q^i$. As for the last part of the statement, the proof of Lemma 3.1 and Lemma 3.2 can be adapted to random variables taking their values over couples $(G;T)$ where $(G;T)$ is a hierarchical clustering, and such that $X_{p,q}(G;T)$ (respectively $Y_{p,q}$) returns the weight of the edge $\{p,q\}$.

## 5 Perspectives and future work

We have established a fundamental and useful property of a modularity measure introduced in [10]. This property being now proved, the $MQ$ measure appears as a prevalent choice over less focused optimization criterion for graph clustering. Indeed, comparing two different clusterings of a same graph as well as asserting the overall quality of a given clustering can now rely on the fact that $MQ$ is gaussian. We however still need to work and provide estimations for the mean and standard deviation of these gaussian approximations.

We have moreover introduced a generalization extending $MQ$ to hierarchical clusterings of graphs. We are now studying the combinatorics of the coefficients appearing in the $MQ(q;\bullet;\bullet)$ expression in order to improve our knowledge on this multilevel modularity measure. Better knowing its mechanics could help us use $MQ$ as a tuning criterion to incrementally compute a "good" hierarchical clustering of a graph. We also need to observe how the mean and standard deviation differ from the usual case.

Although originally defined to work on simple and non-weighted graphs, $MQ$ admits a fuzzy version. Indeed, a fuzzy membership relationship is straightforward to encode in the matrix $\mathbf{C}$ where entries are allowed to be real numbers $c_{i,q} \in [0,1]$ such that $\sum_i c_{i,q} = 1$. In that case, the number $c_{i,q}$ simply reflects the probability that vertex $v_q$ belongs to cluster $C_i$. The quality index $MQ$ must then be adapted to this more general setting by computing the weight $\omega(C_i)$ of clusters $C_i$ (instead of their cardinalities) as $\omega(C_i) = \sum_q c_{i,q}$. Note that this generalization should also require the use of fractional binomial coefficients. Being able to deal with weighted edges in yet another direction we wish to explore.

## References

[1] C. J. Alpert and A. B. Kahng. Recent developments in netlist partitioning: A survey. *Integration: the VLSI Journal*, 19(1-2):1–81, 1995.

[2] D. Auber, Y. Chiricota, F. Jourdan, and G. Melançon. Multiscale navigation of small world networks. In *IEEE Symposium on Information Visualisation*, pages 75–81, Seattle, GA, USA, 2003. IEEE Computer Science Press.

[3] S. Bornholdt and G. Schuster, editors. *Handbook of Graphs and Networks: From the Genome to the Internet*. Wiley-VCH, 2003.

[4] Y. Chiricota, F. Jourdan, and G. Melançon. Software components capture using graph clustering. In *11th IEEE International Workshop on Program Comprehension*, pages 217–226, Portland, Oregon, 2003. IEEE / ACM.

[5] S. N. Dorogovtsev and J. F. F. Mendes. *Evolution of Networks : From Biological Nets to the Internet and WWW*. Oxford University Press, 2003.

[6] W. Feller. *An Introduction to Probability Theory and Its Applications*, volume 2. Wiley, New York, 3rd edition, 1971.

[7] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy Science USA*, 99:7821–7826, 2002.

[8] S. Guha, R. Rastogi, and K. Shim. ROCK: a robust clustering algorithm for categorical attributes. In *Proceedings of the 15th ICDE*, pages 512–521, Sydney, Australia, 1999.

[9] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[10] S. Mancoridis, B. S. Mitchell, C. Rorres, Y. Chen, and E. Gansner. Using automatic clustering to produce high-level system organizations of source code. In *IEEE International Workshop on Program Understanding (IWPC'98)*, Ischia, Italy, 1998.

[11] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physics Reviews E*, 69(026113), 2004.

[12] D. Watts and S. H. Strogatz. Collective dynamics of "small-world" networks. *Nature*, 393:440–442, 1998.

[13] D. J. Watts. *Small Worlds: The Dynamics of Networks between Order and Randomness*. Princeton University Press, 1999.