Journal of Interconnection Networks
Vol. 19, No. 1 (2019) 1940002 (20 pages)
© World Scientific Publishing Company
DOI: 10.1142/S0219265919400024

The Cohesion-Based Communities of Symptoms of the Largest Component of the DSM-IV Network

MOHAMMAD NAZMUL HAQUE

School of Electrical Engineering and Computing,
The University of Newcastle University Drive,
Callaghan, NSW 2308, Australia
Mohammad.Haque@newcastle.edu.au

PABLO MOSCATO*,†

School of Electrical Engineering and Computing,
The University of Newcastle University Drive,
Callaghan, NSW 2308, Australia
Pablo.Moscato@newcastle.edu.au

Accepted 15 March 2019

Modern methods for network analytics provide an opportunity to revisit preconceived notions in the classification of diseases as "clusters of symptoms". Curated collections which were subsequently modified, like the *Diagnostic and Statistical Manuals of Mental Disorders* "DSM-IV" and the most recent addition, DSM-5 allow us to introspect, using the solution provided by modern algorithms, if there exists a consensus between the clusters obtained via a data-driven approach, with the current classifications. In the case of mental disorders, the availability of a follow-up consensus collection (e.g. in this case the DSM-5), potentially allows investigating if the classification of disorders has moved closer (or away) to what a data-driven analytic approach would have unveiled by objectively inferring it from the data of DSM-IV.

In this contribution, we present a new type of mathematical approach based on a global cohesion score which we introduce for the first time for the identification of communities of symptoms. Different from other approaches, this combinatorial optimization method is based on the identification of "triangles" in the network; these triads are the building block of feedback loops that can exist between groups of symptoms. We used a memetic algorithm to obtain a collection of highly connected-cohesive sets of symptoms and we compare the resulting community structure with the classification of disorders present in the DSM-IV.

Keywords: Network analysis; psychopathology; community detection; memetic algorithms; psychometrics.

*http://www.newcastle.edu.au/profile/pablo-moscato.
†Corresponding author.

1. Introduction

Real-world complex systems in biology, physics, economics, marketing, social science, and many other fields are naturally represented as networks or graphs. Edges in these structures represent some sort of interactions, so finding a partition of the set of nodes in groups that have a high number of edges between pairs is an important problem. This has attracted many researchers from the fields mentioned above, generally described as *community identification*.

In psychopathology, some authors have proposed to use techniques from network analytics and graph optimization to elicit the structures present in networks of co-occurrence of symptoms for different disorders. In Ref. 19, authors have extracted the mental disorders and their associated symptoms from the DSM-IV (Diagnostic and Statistical Manual of Mental Disorders, 4th Edition). We can then represent each symptom with a node. An edge, and not an arc, may then be used to connect these symptoms in pairs, with each of the edges indicating that a pair has been seen together in at least one disorder. This network could be useful for further analysis to discover interesting associations of symptoms with disorders.

While we are aware that "correlation does not imply causation", it is pointed by Borsboom and Cramer in Ref. 3 that if the time variable is also considered, some symptoms may have causal connections. Indeed, to illustrate this ideas, in the abstract of their paper[3] they consider the triad *'worry'*, *'insomnia'*, and *'fatigue'*, and use arcs instead of edges to connect them. Assuming these to be causal relationships, it may be plausible that, for some people, 'worry' (about something) may lead to 'insomnia'. Accordingly, over time, the persistence of these two symptoms may finally lead to the 'fatigue' of an individual. Self-stabilization of these may imply the existing of a "feedback loop" mechanism. As implicit in the discussion of Ref. 3, *'substance abuse'* may help lead the individual with the initial causes that make him/her to worry in the first case, thus closing the loop. However, self-stabilization indicates that by measuring correlations/co-occurrence of the symptoms, for some disorders the three symptoms may be present. The occurrence of a triad may then be a characteristic feature in one or more disorders.

Several objective functions have been proposed to help identify these communities of nodes highly interconnected. The use of different mathematical definitions of what constitutes something to be "densely connected" naturally led to different results. One of the most popular global measures used to search for these community structures is called *Modularity*.[15] Over the years, many other approaches have been proposed. There is a genuine interest in the scientific community to understand what they specifically contribute to uncover characteristics of a network's structure.[4]

In this work, we are introducing another new community score and present it in the area of the analysis of symptoms. While we are perfectly aware that this could be perceived as "crowding" an already busy field, one that seems to have several competing proposals seeking for attention. However, we think it is now necessary to introduce a novel alternative metric that uses both weight information and is based on triangles instead of edges. This new measure is perhaps very important as it is based on "triangles", triads like one of the symptoms described before which may be present at the core of disorders in which self-stabilization of symptoms may lead to observing them over prolonged periods of time.

We based our proposal in some previous work, including our own previous definition of the community which is specifically based on identifying triads of edges. In Ref. 13 we have proposed a modification of the *cohesion score* pioneered by Friggeri, Chelius and Fleury and Ref. 10 for the identification of a single community in a network. While other metrics, like modularity, are based on edges, the cohesion score[10] and our modified version[13] are both based on triangles (i.e. cliques of size three).

In Ref. 13 we proposed a memetic algorithm for finding the size of the largest cohesive group; the paper also presented a result that proved the parameterized computational complexity of the natural variants of the problem in Ref. 11. Together, these complexity results indicate that for large instances of the problem, we may need to use sophisticated heuristics and data structures to deal with these type of problems. To address this need, in this paper we aim at extending the previous approach to identify all cohesive groups from the graph, i.e. a partition of the set of nodes (as opposed to identifying a single community as in Ref. 13). In order to identify the communities (based on the cohesion score) we have used memetic algorithm but now informed with an interesting data structure to initialize the population (i.e. a *k-truss decomposition* of the network).

2. The DSM-IV and the DSM-5

The first edition of the *Diagnostic and Statistical Manual of Mental Disorders* was published in 1952 and contained 106 mental disorders.[12] The DSM-I (that started what turned to be the periodical "cycle of classifications"[2]), was followed by the DSM-II in 1968. This new manual went into its Six Edition until work by Robert L. Spitzer and Joseph L. Fleiss showed that it was being used as a diagnostic tool that lacked reliability.[18] Robert L. Spitzer was then selected to chair the committee for the creation of a new manual. The DSM-III was published in 1980 and contained 265 diagnostic categories in 494 pages. The critique on lack of reliability persisted even with a revised version (the DSM-III-R) published in 1987.

Allen Frances, co-author of *"Differential Therapeutics in Psychiatry: The Art and Science of Treatment Selection"*[8] was entitled to chair a Task Force for the creation of the DSM-IV (1994). The new manual included now 410 disorders in its 886 pages. A "text revision" of the DSM-IV (called DSM-IV-TR) was later published. Implicit in the creation of the DSM-IV is the concept of a network of symptoms on which mental disorders are defined: *"there is no assumption that each category of mental disorder is a completely discrete entity with absolute boundaries dividing it from other mental disorders or from no mental disorder."*

After nearly two decades of the DSM-IV being in use, finally, the DSM-5 was published in May 18, 2013. It introduced many changes. For instance, the following five subtypes of schizophrenia (paranoid, disorganized, catatonic, undifferentiated and residual) were removed.[1] Interestingly, one of the most outspoken critics of DSM-5 has been Allen Frances, who judged its introduction *"... the saddest moment in my 45 year career"* and described what are, in his view, the 10 worst changes introduced.[a] He has also written: *"The right goal for DSM-5 would have been diagnostic restraint and deflation, not a further unwarranted expansion of diagnosis and treatment"*.[9] The DSM-5 provides guidelines for 541 diagnoses and several changes.[17] At the time of writing this article, only 5 years after it has appeared in the press, it has been cited 261999 times in the scientific literature. This means that it currently is being cited nearly one time every 10 minutes! The potential for being massively influential for decades is guaranteed. While we do not have its data in network format at this stage, looking at the changes observed in the network from DSM-IV could perhaps help us understand its structure.

3. Basic Mathematical Definitions

To be self-contained, and in order to understand the objective function, we return to some of the formal definitions. Figure 1 shows an illustrative example of a graph of symptoms connected by the co-occurrence in different disorders, with the weight indicating the number of disorders in which both have co-occurred.

3.1. *Cohesion score*

Let $G = (V, E)$ be a simple undirected graph with a set of nodes V and a set of edges E of size $|V| \geq 4$. For a set of nodes $S \subseteq V$ we define an induced subgraph $G[S]$ having a set of edges $E_s \subseteq E$. When we have three nodes connected by three edges we have a *clique* of size three. We can formally define a "triangle in a graph" and two extra definitions that would lead to the concept to 'Cohesion Score' as follows:[13]

Definition 3.1 (Triangle) *A* Triangle *(Δ) in a Graph G is a triplet $(u, v, w) \in V^3$ of pairwise connected nodes, such that $(uv, vw, uw) \in E^3$.*

Definition 3.2 (Inbound Triangle Count) *The* Inbound Triangle Count *for $G[S]$ is denoted as $\Delta_i(G[S]) = |\{(u, v, w) \in S^3 : (uv, vw, uw) \in {E_s}^3\}|$.*

Definition 3.3 (Outbound Triangle Count) *The* Outbound Triangle Count *for $G[S]$ is denoted as the count of triangles which have* exactly two *nodes in S, such that $\Delta_o(G[S]) = \{(u, v, w), (u, v) \in S^2, w \in V \setminus S : (uv, vw, uw) \in E^3|\}$.*

Definition 3.4 (Cohesion Score) *Let $S \subseteq V$ be a subset of nodes in a graph $G = \{E, V\}$, where $\Delta_i(S)$ be the number of triangles with all nodes in the set S and*

[a] https://www.psychologytoday.com/au/blog/dsm5-in-distress/201212/dsm-5-is-guide-not-bible-ignore-its-ten-worst-changes.

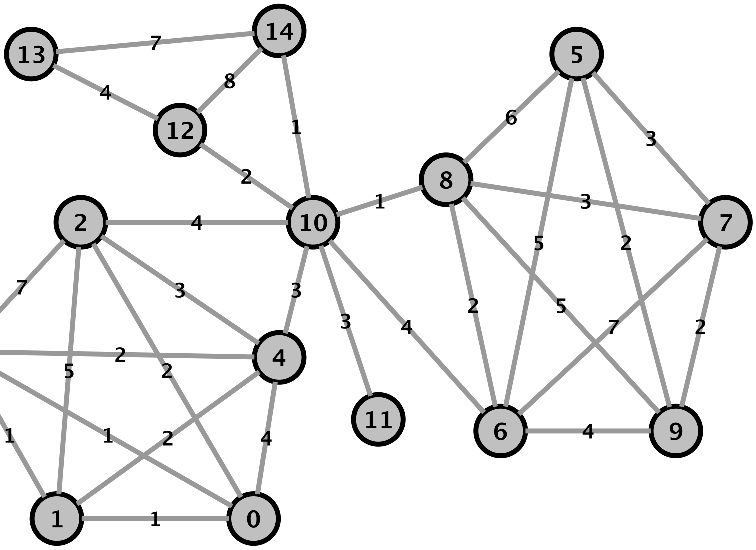


Fig. 1. A weighted graph that will be used to illustrate the community detection approaches in the DSM-IV network. The weight $w(x, y)$ represents the number of co-occurrences of a pair of symptoms x and y. This information will be used to compute a probability, and in turn, these probabilities will influence the results obtained by using one of the objective functions (which we will call *Model 2*). In contrast, *Model 1* does not use these probabilities so it is not based on the number of co-occurrences observed. A single co-occurrence of a pair of symptoms observed in a condition guarantees the presence of an edge in the graph (and they have a weight of 1, as some that can be observed in the figure above).

$\Delta_o(S)$ *the number of triangles with exactly two nodes in* S*, then the* cohesion score of S,[10] *written* $C(S)$*, is defined as:*

$$C(S) = \frac{\Delta_i(S)^2}{\binom{|S|}{3} \times (\Delta_i(S) + \Delta_o(S))} \tag{3.1}$$

3.2. *The k-truss decomposition of a graph*

An important concept used later in our approach can now be defined.

Definition 3.5 (k-truss and k-truss decomposition) *For a given graph* G*, the* k-truss *is the largest subgraph of* G *where every edge is contained in at least* $(k - 2)$ *triangles. Finding all the non-empty* k*-trusses in* G *is referred as the problem of finding a* k-truss decomposition*.*

The *truss number of an edge* $e \in E$ in a graph $G(V, E)$ is the number of triangles in G that contain e. The *k-class of* G is the set of edges that have a truss number equal to k. The outcome of k-truss decomposition of the graph G is shown in Fig. 2. It produces hierarchical subgraphs with different granularity of the cores of the network.[20]

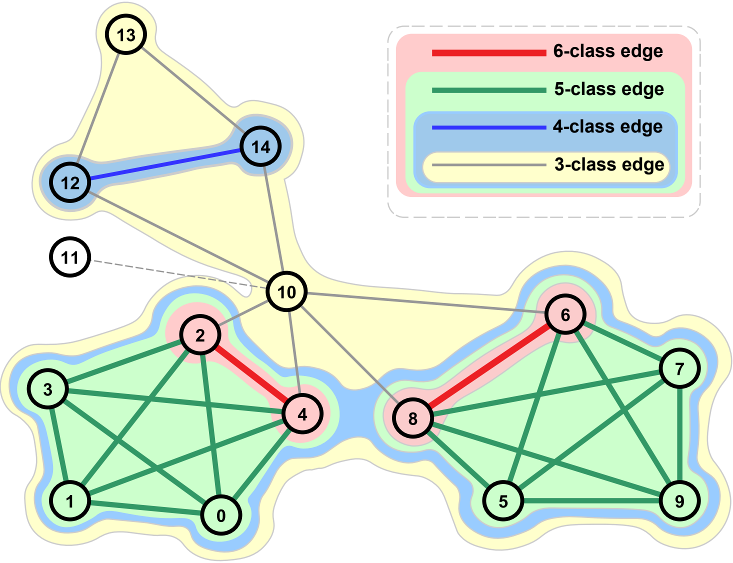


Fig. 2. The k-truss decomposition of the graph shown in Fig. 1 with different layout of nodes and omitted edge weight for clarity. Here, different levels of hierarchies of k-truss are shown in the graph using different colors of edges. Corresponding k-truss communities are also marked with different colors.

4. Objective Functions

4.1. *Modularity*

Modularity[15] is a quality measure for detecting communities in graphs and it is based on edges and does not take into consideration the presence of triangles. If m is the number of edges in the graph and k is the number of communities (the number of sets in the partition P of the nodes of the graph), we can then define the modularity of graph as:

$$Q = \sum_{p=1}^{k=|P|} \left[\frac{m_p}{m} - \left(\frac{D_p}{2m} \right)^2 \right] \tag{4.1}$$

where m_p denotes the number of edges connecting nodes in community p and D_p denotes the total sum of the degrees of all nodes in community p.

The *Modularity* score is based on the ratio of the number of edges (i.e. $\frac{m_p}{m}$) in each of the communities. It also considers the degrees ratio of nodes in the group over the graph (i.e. $\frac{D_p}{2m}$).

4.2. *Connected Cohesion-based objective function for multi-community detection*

Unlike modularity, the connected-cohesion score for a single community S, denoted as $C(S)$, is calculated using the number of triangles (in-bound Δ_i(S) and out-bound Δ_o(S)) for the group S in the graph G. The identification of a set of nodes S^* that

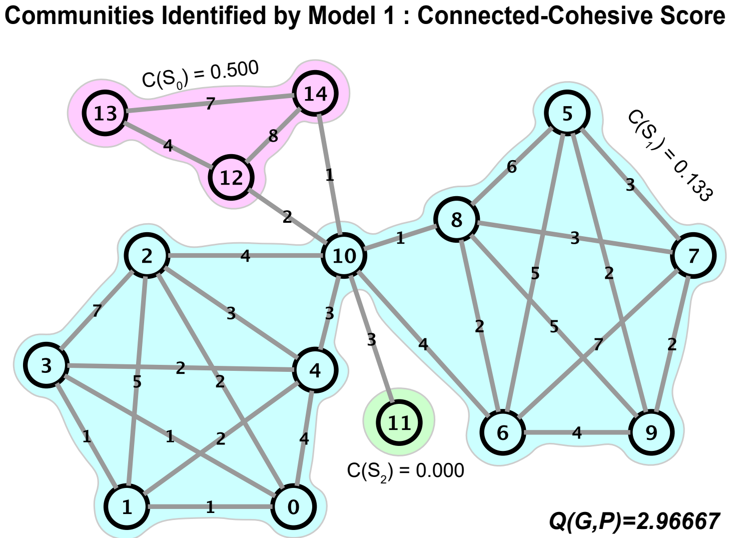


Fig. 3. An example of the Communities identified by optimization of the Connected Cohesion-score. We call this the result of *Model 1*. We have identified three communities $P = \{S_0, S_1, S_2\}$ with Connected Cohesion-score of $C(S_0) = 0.500$, $C(S_1) = 0.133$ and $C(S_2) = 0.000$. The global connected-cohesion score of the graph is $Q(G, P) = 2.96667$. We note that the weights on the node have no role in the calculation of this score.

maximizes this score reveals a group that has many inbound triangles and a small number of outbound triangles.

We build on this connected-cohesion definition as the foundation stone of a new global objective function that would replace the in-group edge ratio of modularity score calculation and permits the identification of multiple communities in the network, thus balancing inbound triangles and outbound triangles in a global way.

Consider now a graph $G(V, E)$, and let $P = \{S_1, S_2, \ldots S_k\}$ be a partition of the set of nodes of V. Then we define the *global connected-cohesion score of the partition* P *on a graph* G as:

$$Q(G, P) = \sum_{i=1}^{k} \left[|S_i| \times C(S_i)\right]. \qquad (4.2)$$

Figure 3 shows an example of the communities identified by the fitness function (given by Eq. (4.2)), the results on graph G. The Connected Cohesion-based objective function identified four communities, $P = \{S_0, S_1, S_2\}$ in G. The community $S_0 = \{12, 13, 14\}$ has one in-bound triangle $(12, 13, 14)$ and one out-bound triangle $(10, 12, 14)$. The Connected Cohesion-score of this community is $C(S_0) = 0.5$. The second community $S_1 = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ has 22 in-bound triangles and no out-bound triangle. The Connected Cohesion-score of this community is $C(S_1) = 0.133$. The third community S_2 contains a single node 11, hence the Connected Cohesion-score is $C(S_1) = 0$. The *global connected-cohesion* score of the partition P on the graph G is $Q(G, P) = 2.96667$.

4.3. *Global Connected-Cohesive Surprisal*

We will also define another objective function which is motivated by this study. Let $n(x, y)$ be the number of times that we have observed the co-occurrence of symptoms x and y in the set of disorders. Let $p(x, y)$ be the ratio $n(x, y)/\sum n(x, y)$. Obviously, the sum of all these ratios equals one, and we will interpret them as an a priori probability of observing such a co-occurrence in the database (and denote it as $p(x, y)$. We can then associate the *surprisal of a co-occurrence of symptoms* as $s(x, y) = log_2(1/p(x, y))$. We can now define the *connected cohesive surprisal of a set of nodes* as:

$$C'(S) = \frac{\Delta_i'(S)^2}{\binom{|S|}{3} \times (\Delta_i'(S) + \Delta_o'(S))} \tag{4.3}$$

where $\Delta_i'(S)$ is the total sum of surprisals of each inward triangle in S, and analogously, $\Delta_o'(S)$ is the total sum of surprisals of outward triangles. The surprisal of a triangle, i.e. the surprisal of three symptoms x, y, z that are observed to co-occur in at least one disorder (not necessary in the same ones), is denoted as $s(x, y, z)$ and it is given by

$$s(x, y, z) = log_2(1/p(x, y)) + log_2(1/p(y, z)) + log_2(1/p(z, x)), \tag{4.4}$$

so

$$\Delta_i'(S) = \sum_{x,y,z} s(x, y, z), \forall x, y, z \in S \tag{4.5}$$

and

$$\Delta_o'(S) = \sum_{x,y,z} s(x, y, z), \forall x, y \in S, z \notin S. \tag{4.6}$$

We deliberately apostrophe here only to highlight the fact that Eq. (3.4) originally inspired by the work of Friggeri, Chelius and Fleury[10] and are similar in intent, but ours accounts for the sum of surprisals of triangles (both inbound and outbound) (in Eq. (4.4)) instead of just counting for how many of them are.

Accordingly, the *global connected cohesion surprisal* of a partition given a graph $G(V, E)$ and a partition of its nodes P is given by:

$$Q'(G, P) = \sum_{i=1}^{k} \left[|S_i| \times C'(S_i)\right]. \tag{4.7}$$

The key modeling idea here is that of finding highly connected sets of symptoms that have co-occurred, such that there is evidence of a possible triad involvement, and that also are less frequent. Finding a partition of the set of nodes such that it maximizes the Q', when using the network of the DSM-IV, help us to identify "groups of symptoms" which are co-observed and yet relatively rare to be jointly observed. In Fig. 4 we observe that now symptom 10 is now "clustered" with the group containing symptoms 13, 12 and 14. While symptoms 10 has been observed as

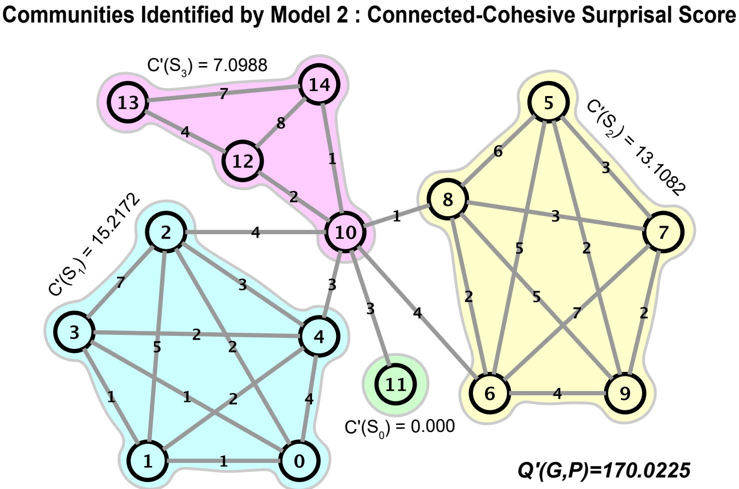


Fig. 4. An example of the communities identified by Model 2, based on the Connected-Cohesive Surprisal score. This model identified four communities $P = \{S_0, S_1, S_2, S_3\}$ with connected-cohesion surprisal score of $C'(S_0) = 0$, $C'(S_1) = 15.217$, $C'(S_2) = 13.108$ and $C'(S_3) = 7.099$. The global connected-cohesion surprisal of the graph G for the collection of four communities P is $Q'(G, P) = 170.0225$.

co-occurring with other symptoms, this objective function "recommends" it being associated to the group, increasing its specificity.

We will use this score as the fitness function for the memetic algorithm. It will only consider the Connected-Cohesive groups to calculate the score which are connected to at least another group. Any disconnected nodes or groups or node which does not contribute any triangle will not be included in the fitness score calculation process. We will then aim at correlating the observations of these groups with the ground truth (i.e. the labels of disorders in the DSM-IV). The method is, however, quite general and can be used in other data mining settings.

Table 1. A snapshot of the values require to compute the Surprisal score for the communities of graph shown in Fig. 1.

x	y	n(x,y)	p(x,y)	1/p(x,y)	s(x,y)
0	2	2	0.019230769	52	5.700439718
0	1	1	0.009615385	104	6.700439718
...	...	...	...	...	...
10	11	3	0.028846154	34.66666667	5.115477217
10	12	2	0.019230769	52	5.700439718
10	14	1	0.009615385	104	6.700439718
12	14	8	0.076923077	13	3.700439718
12	13	4	0.038461538	26	4.700439718
13	14	7	0.067307692	14.85714286	3.893084796
$\sum \mathbf{n(x,y)}$	**104**				

4.4. *An illustrative example of the calculation*

Table 1 displays a subset of the symptoms and the co-occurrence of symptoms in different disorders from the illustrative graph in Fig. 1. The weights of the edges are shown in the column $n(x,y)$. The total sum of the weights of all the edges $(\sum_{x\neq y} n(x,y))$ in the graph is 104. The ratios $(n(x,y)/\sum n(x,y))$ for each of the edges are shown in the column labeled $p(x,y)$. The column $s(x,y)$ shows the values of the surprisals of a co-occurrence of symptoms and it is given by $log_2(1/p(x,y))$.

Let us consider the community $S_3 = \{10, 12, 13, 14\}$ of Fig. 4 to compute the contribution to the Global Connected-Cohesive Surprisal score using only these four nodes and the values of Table 1. To calculate the Connected-Cohesive Surprisal score we need to compute the total sum of surprisals of each triangle (both of the inwards and outwards) associated with the community. The community S_3 consist of four nodes and has only two inward triangles ((10, 12, 14) and (12, 13, 14)). The surprisal score of those triangles can be calculated according to the Eq. (4.4) as:

$$\begin{aligned} s(10,12,14) &= s(10,12) + s(10,14) + s(12,14) \\ &= 5.700439718 + 6.700439718 + 3.700439718 \\ &= 16.10131915 \\ s(12,13,14) &= s(12,13) + s(12,14) + s(10,14) \\ &= 4.700439718 + 3.700439718 + 3.8930847968 \\ &= 12.29396423 \end{aligned}$$

Now we calculate the connected cohesive surprisal of community S_3 using Eq. (4.3) as:

$$\begin{aligned} \Delta'_i(S) &= s(10,12,14) + s(12,13,14) \\ &= 16.10131915 + 12.29396423 \\ &= 28.39528339 \\ \Delta'_o(S) &= 0.00 \end{aligned}$$

$$\begin{aligned} C'(S) &= \frac{\Delta'_i(S)^2}{\binom{|S|}{3} \times (\Delta'_i(S) + \Delta'_o(S))} \\ &= \frac{28.3952833^2}{\binom{4}{3} \times (28.3952833 + 0.00)} \\ &= \frac{28.3952833^2}{4 \times (28.3952833)} \\ &= 7.098820847 \end{aligned}$$

Hence, the connected cohesive surprisal score for the community S_3 is 7.0988.

5. The Memetic Algorithm for Community Detection

We employ a memetic algorithm to find the communities in the graph. The memetic algorithm follows the genetic algorithm's paradigm of recombination, mutation and the survival of fittest. We used a local-search to improve the individuals. The memetic algorithm is implemented with the help of the DEAP evolutionary computation framework.[7] We used the fitness function in Eq. (4.2) and Eq. (4.7) to evaluate the individuals.

5.1. *Individual representation*

We used the integer array of the lengths equals to the number of nodes in the graph to represent the individual used in the memetic algorithm. Here, each i-th array value in the individual represents the community label of the i-th node.

5.2. *Population initialization*

The initialization of the population is an important step in evolutionary algorithms since it could dramatically affect the convergence and generally helps to improve the quality of the final solutions obtained.[16] However, following the work of other researchers like Krasnogor et al.[14] we generate a small number of individuals in the initial population with problem-specific information.

Our approach is based on computing the k-truss decomposition of the input graph[6] (see Fig. 2 for an example). To initialize the population, we first start by computing the k-truss decomposition of the graph G (as described in Sec. 3.2). It produces a hierarchy of subgraphs.[20] Individuals generated with the information from different truss levels of the graph will provide good diversity to the population of the memetic algorithm (in Fig. 2, we have four edge class values: 3, 4, 5 and 6). Each layer of the hierarchy is represented as an initial individual of the population.

At first we assign label 0 in the individual as the *community label* for all nodes in the graph. Then we update the *community label* of each node with the value of k for each of the incident nodes of edges in the k-truss subgraph. For an example, the

Table 2. Initial subpopulation generated taking into account the k-truss decomposition of the graph. The example follows the graph shown in Fig. 2. Since the maximum k is four, there are four individuals created this way. The nodes connected by the edges in corresponding k-truss subgraphs are being assigned with the same value of k (as the community label) and the remaining nodes received 0 as their community label.

	initial solutions generated from k-truss decomposition of graph																
k-class edge	*node:*	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	
3	[	3	3	3	3	3	3	3	3	3	3	3	0	3	3	3	]
4	[	4	4	4	4	4	4	4	4	4	4	0	0	4	0	4	]
5	[	5	5	5	5	5	5	5	5	5	5	0	0	0	0	0	]
6	[	0	0	6	0	6	0	6	0	6	0	0	0	0	0	0	]

6-truss subgraph of G in Fig. 2 has two edges: $(2, 4)$ and $(8, 6)$. The incident nodes of those edges are $\{2, 4, 6, 8\}$. We assign the *community label* for corresponding indices of those nodes with 6, the value of k in the hierarchy. Hence, the initial individual representing the 6-truss is [0, 0, 6, 0, 6, 0, 6, 0, 6, 0, 0, 0, 0, 0, 0] for sequential ordering of the node ids (0 to 14) in G.

All the initial individuals for k-truss decomposition of the graph shown in Fig. 2 are presented in Table 2. The rest of the individuals in the population are generated by selecting uniformly at random a label which is an integer in the set $\{1, \ldots, |V|\}$, where $|V|$ is the number of nodes in the graph.

5.3. *Genetic operators*

To generate new individuals for the next generation, we used two-point crossover.[b] This could result in highly-disruptive new solutions but an iterative improvement procedure compensate this and is explained next.

5.4. *Improvement of solutions using a local search-based heuristic*

We used local-search to improve the quality of solutions. After a certain interval of generations of the Memetic Algorithm, a local search-based improvement is applied. For all communities in a solution, an improvement heuristic is based on the execution of four steps. In the first step, for a community in the individual (i.e. a set of nodes that share the same label), we search for one neighboring node which can create an inbound triangle (i.e. it had already two nodes in that community). The same label is then assigned to that node. We note that this first step does not check for the change in the global objective function. Next, for all the nodes in the community, we check if they are not contributing to inbound triangles. If they do not contribute, then a new label not yet used in for this solution is assigned to them (i.e. if the maximum label currently for the individual is k', then it will be assigned the value $k' + 1$). In the third step, we recalculate the objective function value. Finally, we seek to expand the community by adding new nodes to it. This step now checks for the global fitness function when doing the updates. We include a new node to the community unless the additional node decreases the objective function value for the individual. The detailed description of the process is shown at Algorithm 1.

6. Computational Results using DSM-IV Data

6.1. *The largest subnetwork of the DSM-IV*

To produce the DSM-IV graph, we used the information available on the public domain[c] facilitated by the authors of the paper *"Mapping the manuals of madness:*

[b] deap.tools.cxTwoPoin found at: http://deap.readthedocs.io/en/master/api/tools.html.
[c] https://sites.google.com/site/dsmgraphs/Home/files.

Algorithm 1: The INDIVIDUAL SEARCH optimization

```
Input: A solution $S$ and the Graph $G$
Output: The improved solution $S_{opt}$
/* Initialisation of variables */
$Improve \leftarrow True$
$Nodes_{out} \leftarrow G$.get$\Delta$Nodes() $\setminus$ $S$.getNodes()
/* add a $\Delta$neighbour */
foreach $\{(u, v, w) \in Nodes_{out}$ do
    if is$\Delta(u, v, w) = True$ then
        $S$.extend($u, v, w$)
        break
    end
end
/* Prune: Remove nodes not in $\Delta_i$ */
foreach $v \in S$ do
    if $\Delta_i(v,\ S) < 1$ then
        $Nodes_{rem}$.append($v$)
    end
end
$S_{opt} \leftarrow S \setminus Nodes_{rem}$
/* Expand Neighbourhood: add more $\Delta_o$(having 2 in group neighbours) */
while $Improve$ is $True$ do
    $Improve \leftarrow False$
    $Nodes_{avail} \leftarrow G$.get$\Delta$Nodes() $\setminus$ $S_{opt}$.getNodes()
    foreach $v \in Nodes_{avail}$ do
        if $\Delta_o(v,\ S) < 1$ then
            $S_{opt}$.append($v$)
            if $S_{opt}$.fitness() improves then
                $Improve \leftarrow True$
            else
                $S_{opt}$.remove($v$)
            end
        end
    end
end
return $S_{opt}$
```

Comparing the ICD-10 and DSM-IV-TR using a network approach".[19] Using the DSM-IV information, a network of 439 nodes (one for each of the symptoms) was created. An edge in the network exists between two nodes if the two symptoms that the edge connects have co-occurred for at least one disorder (from a set of 148 disorders). The resultant graph consists of 439 nodes (symptoms) and 2626 edges.

The DSM-IV graph has a giant component (the *largest connected component* or the *core* of a graph) consisting of 208 nodes and 1949 edges. This is the object of this study. We thus created a weighted graph for this giant component by adding an edge weight, the number of times a pair of symptoms co-occurred for the same disorder. We used this weighted graph for the experiments and we have computed the surprisals as described in Sec. 4.3.

6.2. *The communities of the largest subnetwork of DSM-IV*

We repeated the execution of the algorithms (both of the *Model 1* and *Model 2*) for 100 independent runs on DSM-IV's core graph. The best solution found by each of the models were compared for the inter-rater agreement of the separation of nodes in different communities. We have used the Cohen's Kappa score[5] to measures the inter-rater agreement of the community labelling. The agreement score for the community labeling of nodes between the result of *Model 1* and *Model 2* is 0.777. According to the accepted rules for its interpretation, it represents that a good, but not perfect, agreement exists between two models. This is good news as it indicates that the "surprisal" based model is indeed different than the other one and we will discuss some of the differences found. The summary of the fitness scores achieved for the execution is shown in Table 3. The communities identified in DSM-IV core graph (yEd software's[d] organic layout with natural clustering quality 0.7 is used to generate the visuals) for the popular Modularity score (described in Sec. 4.1) and *Model 1* are shown in Fig. 5(a) and Fig. 5(b), respectively.

Table 3. Summary of the results for 100 repeated independent execution of the experiments on DSM-IV.

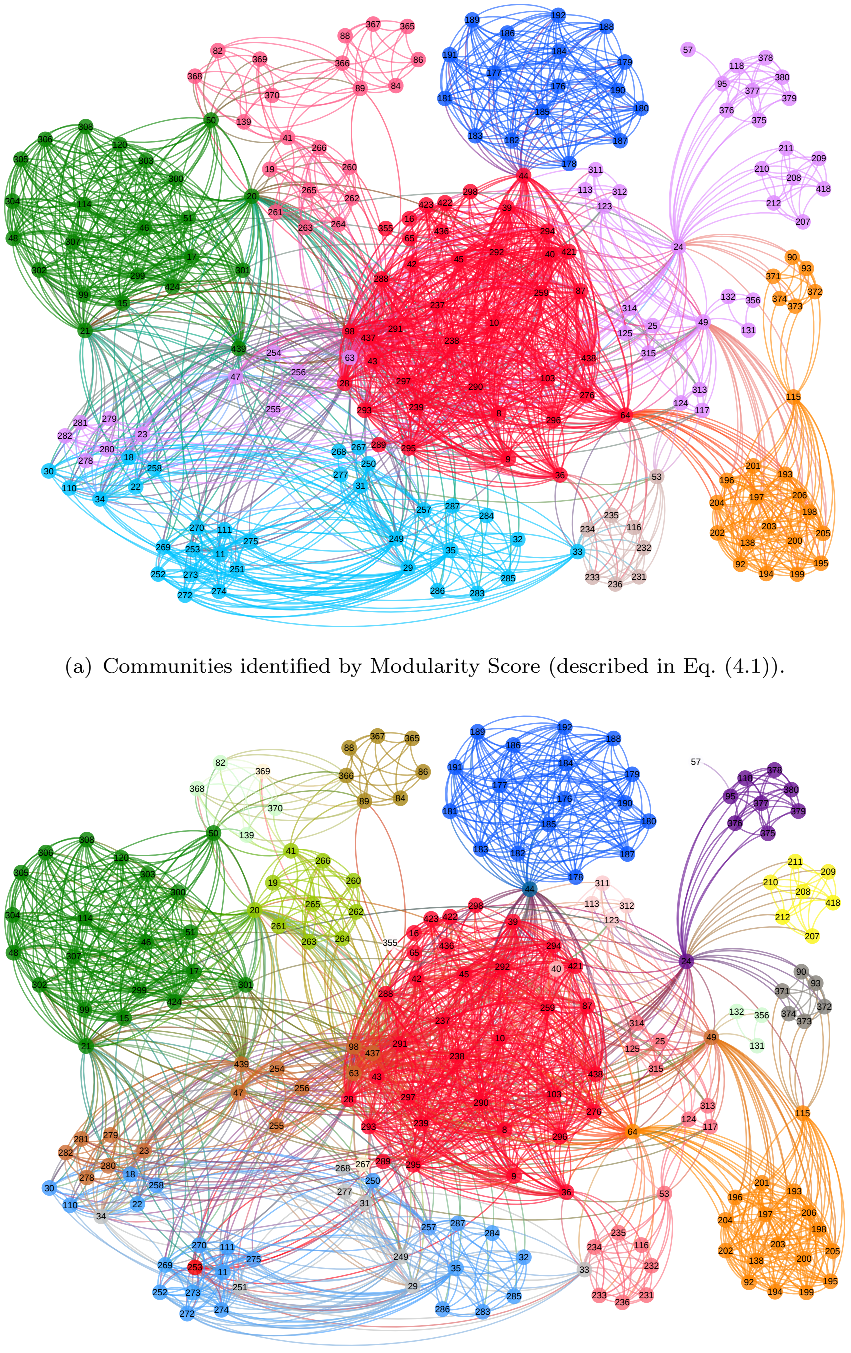
	Fitness Score			
Experiment	(Max)	(Avg. ± Std. Dev.)	(Min)	Kappa Score
Model 1	93.1198	61.9792 ± 5.0857	42.3078	0.7771
Model 2	4102.4889	3279.0420 ± 417.5231	1544.2162	

The Communities detected in the core of DSM-IV by the Connected-Cohesive surprisal score based Model 2 is shown in Fig. 6. We used different colors to show the communities, identifier numbers to mark and labels to distinguish them. To create community labels, we counted the frequencies for each of the disorders in each of the community and brought together the names of the most dominant disorders.

6.3. *Some findings of interest*

There are some symptoms worth discussing as the different models gave different memberships. For instance, symptom *44, "often easily distracted by extraneous stimuli / distractibility"*, is located, both for *Model 1* and *Model 2* with those of Attention Deficit / Hyperactivity Disorder. Symptom *40, "difficulty concentrating / diminished ability to think or concentrate, or indecisiveness / poor concentration or difficulty making decisions / difficulty concentrating or mind going blank"*, is grouped differently only in *Model 1.* Symptoms *33, "memory impairment / impairment in attention or memory"*, is in a community for *C15, Vascular Dementia and Dementia of the Alzheimer's Type*, only for the surprisal-based *Model 2* and *253, "Stupor"*

[d] https://www.yworks.com/products/yed.



(a) Communities identified by Modularity Score (described in Eq. (4.1)).

(b) Communities identified by MA using the objective function for *Model 1*.

Fig. 5. The communities identified in DSM-IV core graph using a) *Modularity* score and b) *Model 1*, the Connected-Cohesion score. Clearly, *Modularity*, as an edge-based metric, tends to "merge" in the same community several densely connected subnetworks, an aspect that may be less desirable to separate groups of symptoms, making the communities obtained by *Modularity* less specific.

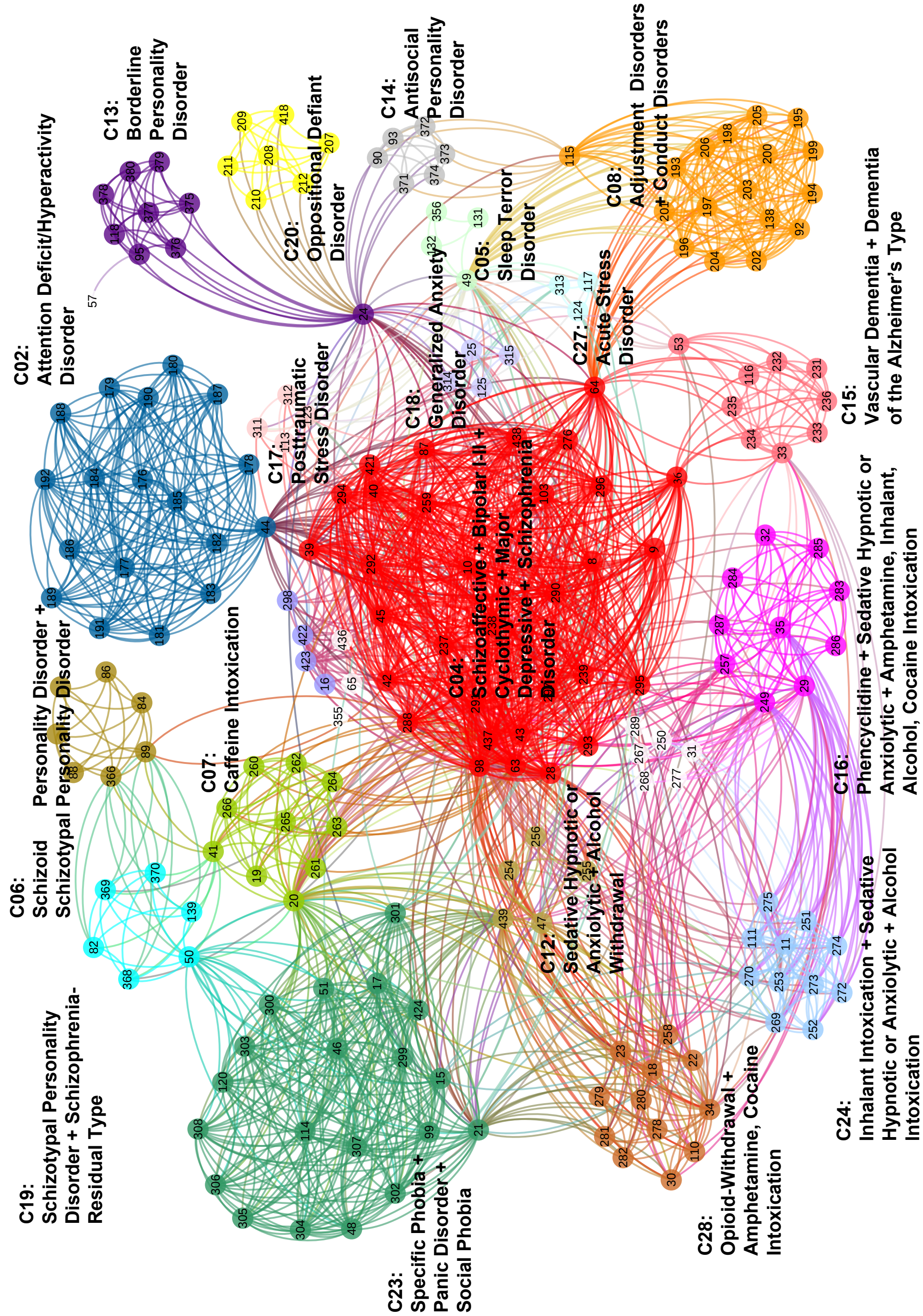

C02: Attention Deficit/Hyperactivity Disorder
C13: Borderline Personality Disorder
C20: Oppositional Defiant Disorder
C14: Antisocial Personality Disorder
C08: Adjustment Disorders + Conduct Disorders
C05: Sleep Terror Disorder
C18: Generalized Anxiety Disorder
C27: Acute Stress Disorder
C17: Posttraumatic Stress Disorder
C15: Vascular Dementia + Dementia of the Alzheimer's Type
C04: Schizoaffective + Bipolar I-II + Cyclothymic + Major Depressive + Schizophrenia Disorder
C16: Phencyclidine + Sedative Hypnotic or Anxiolytic + Amphetamine, Inhalant, Alcohol, Cocaine Intoxication
C06: Schizoid Personality Disorder + Schizotypal Personality Disorder
C07: Caffeine Intoxication
C12: Sedative Hypnotic or Anxiolytic + Alcohol Withdrawal
C19: Schizotypal Personality Disorder + Schizophrenia-Residual Type
C23: Specific Phobia + Panic Disorder + Social Phobia
C28: Opioid-Withdrawal + Amphetamine, Cocaine Intoxication
C24: Inhalant Intoxication + Sedative Hypnotic or Anxiolytic + Alcohol Intoxication

Fig. 6. Communities obtained by the memetic algorithm for the core subnetwork of the DSM-IV when a surprisal-based weighting is used (i.e. *Model 2* is used as the objective function). The communities are distinguished by different colors and tentatively annotated using the most prevailing disorders to which the symptoms have been associated.

is within the Inhalant, Sedative Hypnotic or Anxiolytic, and Alcohol Intoxication group in *Model 2*. This model also separates well these symptoms to other types of intoxications and withdrawals. Symptoms *439, "Nausea"* is within the *C12, "Sedative Hypnotic or Anxiolytic and Alcohol Withdrawal"* community found by *Model 2*, and that symptom *34, "Pupillary dilation"* is within community *C28, "Opioid-Withdrawal and Amphetamine and Cocaine Intoxications"*. This community comes very clearly by the use of the surprisal-based *Model 2* but not so coherently by the other two objective functions *Model 1* and *Modularity*. Finally symptom *64, "Depressed mood*, appears in the largest cluster of Fig. 6 (center, in red), for *Model 2* and not as part of the group that is labeled as *C08, Adjustment Disorders or Conduct Disorders*.

6.4. *Substance-related disorders*

Our algorithm has identified a total of five communities with symptoms related to *"Substance related disorders"* shown in Fig. 7. Those communities are: *C07, Caffeine Intoxication*, *C12, Sedative Hypnotic or Anxiolytic Withdrawal*, *C28, Opioid Withdrawal*, *C24, Inhalant Intoxication* and *C16 : Phencyclidine Intoxication*.

Symptom *20, "Tachycardia / Accelerated heart rate"* is an articulation point in this subnetwork (i.e. its removal would render this subnetwork separated in two connected components). However, inspection of Fig. 6 shows it also co-occurring with other symptoms different from those related to substance use. However, the use of

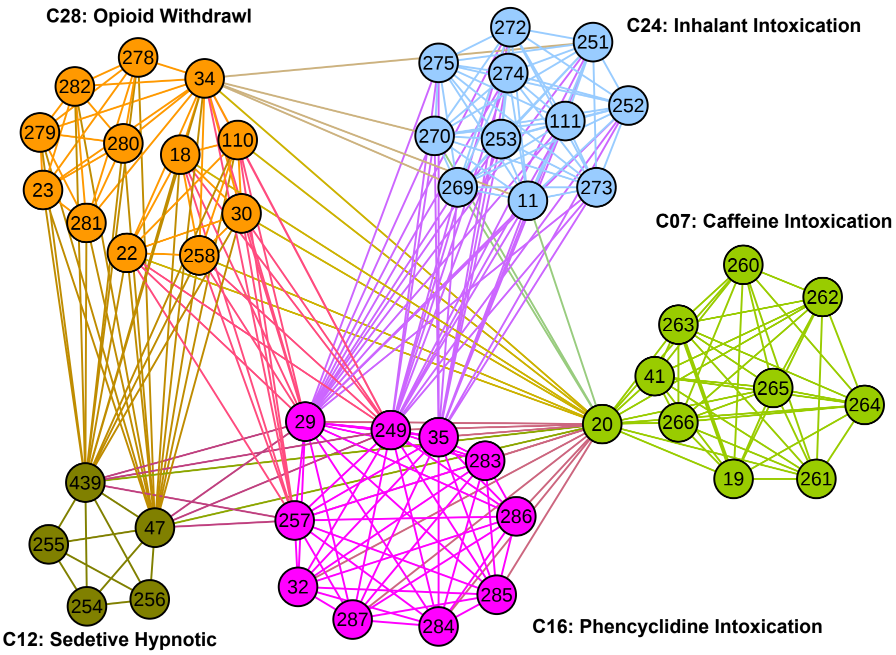


Fig. 7. Communities identified related to Substance-related disorders. Here we only show the edges which are within the community and connect only these communities.

the surprisal and triangle-based approach that the *Model 2* embodies "recommends" its inclusion within the *C07, Caffeine Intoxication*-rich group of symptoms.

From Fig. 7 we can see that two symptoms (*47, Vomiting* and *439, Nausea*) in a community for *C12, Sedative hypnotic or Anxiolytic withdrawal* has the most co-occurrences with the symptoms from *C16, Phencyclidine Intoxication* and *C28, Opioid Withdrawal.* On the other hand, symptom *34, Pupillary dilation* is in the community of a majority of symptoms related to *C28, Opioid withdrawal* and is the only node having the connection with the symptoms from the group of *C24, Inhalant intoxication* disorders.

Three symptoms (*257, Elevated blood pressure / hypertension*, *29, Coma* and *249, Clinically significant maladaptive behavioral or psychological changes that developed during or shortly after substance ingestion*) from the disorder group of *C16, Phencyclidine Intoxication* has all of the co-occurrences of disorders from the *C28, Opioid Withdrawal* group. Two symptoms (29, 24) form those and *35: Nystagmus / vertical or horizontal nystagmus* in community C16 also have a similar sort of relation with *C24, Inhalant Intoxication* disorders.

6.5. ***Symptoms in a community of their own***

The results obtained with *Model 2* reveal several communities having a single symptom. There are two main reasons for this. In some cases, the symptoms have co-occurred with others but no other symptom has co-occurred with both. As a consequence, they are excluded from another community and form a community of their own. Two cases exist: *57, "Unexpected travel away from home or one's customary place of work, with inability to recall one's past"* and *355, "On awakening from the frightening dreams, the person rapidly becomes oriented and alert"*.

Others could have been members of a minimum of four and a maximum of eight communities. On example is symptom *24, "is often touchy or easily annoyed by others / irritability, frustration, or anger / irritability or outbursts of anger / irritability and aggressiveness / inappropriate, intense anger or difficulty controlling anger* which has been grouped with others of *C13, Borderline Personality Disorder.* Instead, they have been not clustered with anyone of these. The reason can be investigated further by a closer look at the connections of node 24 with other communities (*C13*, *C20*, *C14*, *C5*, *C18* are in descending order of their $|S_i| \times C'(S_i)$ value). Among those communities, symptom *24* creates the largest clique of size 9 with the *C13.* Next two candidate communities it could have formed are with C20 (clique size 8) and C14 (size 7 clique), respectively. However, a community with the biggest clique will contain the most number of inbound triangles, which will positively contribute to the fitness score. Therefore, the algorithm has chosen the symptom 24 to include in the community *C13*, than in other candidate communities.

Another symptom, *49, "Anxiety"*, shared by different disorders, has clustered in a group of other three symptoms which we have generically labeled under *C5, Sleep Terror Disorder.* Other singled out of other communities are: *355, "On* *awakening from the frightening dreams, the person rapidly becomes oriented and alert"*; *267, "Conjunctival injection"*; *289, "Catalepsy"*; *250, "Recent ingestion of alcohol"*; *65, "Vivid, unpleasant dreams / Repeated awakenings from the major sleep period or naps with detailed recall of extended and extremely frightening dreams"*; *31, "Drowsiness"*; *277, "Pupillary constriction"*; *268, "Dry mouth"*; and *436, "Dysphoric mood"*.

7. Conclusion

In this work, we have presented the extension of the work[13] where the memetic algorithm was used for community detection using the cohesion score to find a single community. Here, we used two cohesion scores to quantify the natural community behavior of the communities present in complex networks. The memetic algorithm is initialized with the intelligent k-truss based population initialization mechanism which helped to generate diversity in the population. The multiple communities have stronger triangle-based interconnection and less outside the group connection which able to find interesting insights from the network.

We note, once again, that the disorders have never been used to inform the creation of the communities, so we conclude that the *Model 2* brings groups of symptoms which highly correlates with the *"Disorder class"* annotation of DSM-IV. It also brings important information that could have guided the development of the DSM 5. The use of this method may have helped the Task Force working in the DSM 5 to better characterize the modular structure of the large dataset on symptoms observed over many disorders in the order to carefully modify the network towards better characterizations, improving the sensitivity and specificity if used for research and diagnosis purposes.

Acknowledgments

We thank the timely help from Pia Tio and Denny Borsboom from the Department of Psychological Methods, University of Amsterdam, that facilitated the DSM-IV network and Ademir Gabardo for help in the preparation of the weighted network. P. M. acknowledges previous support from the Australian Research Council Future Fellowship FT120100060 and Australian Research Council Discovery Projects DP120102576 and DP140104183.

References

1. American Psychiatric Association, "Highlights of Changes from DSM-IV to DSM-5," *FOCUS* **11**(4) (2013), 525–527. https://doi.org/10.1176/appi.focus.11.4.525.
2. R. K. Blashfield, J. W. Keeley, E. H. Flanagan, and S. R. Miles, "The cycle of classification: DSM-I through DSM-5," *Annual Review of Clinical Psychology* **10** (2014), 25–51.
3. D. Borsboom and A. O. Cramer, "Network Analysis: An Integrative Approach to the Structure of Psychopathology," *Annu Rev Clin Psychol.* **9** (2013), 91–121.

4. C. Cheng-Shang, L. Duan-Shin, L. Li-Heng, L. Sheng-Min, and W. Mu-Huan, "A Probabilistic Framework for Structural Analysis and Community Detection in Directed Networks," *IEEE/ACM Trans. Netw.* **26** (2018), 31–46.
5. J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement* **20** (1960), 37–46.
6. J. Cohen, "Trusses: Cohesive subgraphs for social network analysis," *National Security Agency Technical Report*, **16** (2008).
7. F.-A. Fortin, F.-M. D. Rainville, M.-A. Gardner, M. Parizeau, and C. Gagné, "DEAP: Evolutionary Algorithms Made Easy," *Journal of Machine Learning Research* **13** (2012), 2171–2175.
8. A. Frances, *Differential therapeutics in psychiatry: the art and science of treatment selection*, Brunner-Routledge, 1984.
9. A. Frances, *Saving normal: An insiders revolt against out-of-control psychiatric diagnosis, DSM-5, big pharma and the medicalization of ordinary life,* William Morrow, 2013.
10. A. Friggeri, G. Chelius, and E. Fleury, "Triangles to Capture Social Cohesion," *Proc. 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing,* IEEE, 2011, pp. 258–265.
11. A. Friggeri, and E. Fleury, "Maximizing the Cohesion is NP-hard," https://arxiv.org/pdf/1109.1994, 2011.
12. G. N. Grob, "Origins of DSM-I: A study in appearance and reality," *The American Journal of Psychiatry* **148** (1991), 421–431.
13. M. N. Haque, L. Mathieson, and P. Moscato, "A memetic algorithm for community detection by maximising the connected cohesion," *Proc. 2017 IEEE Symposium Series on Computational Intelligence (SSCI),* IEEE, 2017, pp. 1–8.
14. N. Krasnogor, A. Aragón, and J. Pacheco, "Memetic algorithms," *Metaheuristic Procedures for Training Neutral Networks,* Springer, 2006, pp. 225–248.
15. M. E. J. Newman, and M. Girvan," Finding and evaluating community structure in networks," *Phys Rev E Stat Nonlin Soft Matter Phys,* **69** (2004), 1–15.
16. S. Rahnamayan, H. R. Tizhoosh, and M. M. Salama, "A novel population initialization method for accelerating evolutionary algorithms," *Computers & Mathematics with Applications,* **53** (2007), 1605–1614.
17. D. A. Regier, E. A. Kuhl, and D. J. Kupfer, "The DSM-5: Classification and criteria changes," *World Psychiatry,* **12** (2013), 92–98.
18. R. L. Spitzer and J. L. Fleiss, "A Re-analysis of the Reliability of Psychiatric Diagnosis," *British Journal of Psychiatry,* **125** (1974), 341–347.
19. P. Tio, S. Epskamp, A. Noordhof, and D. Borsboom, "Mapping the manuals of madness: comparing the ICD-10 and DSM-IV-TR using a network approach," *International journal of methods in psychiatric research,* **25** (2016), 267–276.
20. J. Wang and J. Cheng, "Truss decomposition in massive networks," *Proc. of the VLDB Endowment* **5** (2012), 812–823.