Community landscapes: an integrative approach to determine overlapping network module hierarchy, identify key nodes and predict network dynamics

I. A. Kovács, R. Palotai, M. S. Szalay and P. Csermely*
Department of Medical Chemistry, Semmelweis University, Tűzoltó str. 37-47, H-1094 Budapest, Hungary
*e-mail: csermely@eok.sote.hu

Network communities help the functional organization and evolution of complex networks. However, the development of a method, which is both fast and accurate, provides modular overlaps and partitions of a heterogeneous network, was rather difficult[1-7]. Here we introduce the novel concept of community landscapes and ModuLand, an integrative framework determining overlapping network modules as hills of the community landscape and including several widely used modularization methods as special cases[1,3]. The current 3+2 implementations of the community landscape concept provide a fast analysis of weighted and directed networks, and (1) determine overlapping modules with a high resolution; (2) uncover a hierarchical network structure in previously unprecedented details allowing a fast, zoom-in analysis of large networks; (3) allow the determination of key network elements and (4) help to predict network dynamics. The concept opens a wide range of possibilities to develop new approaches and applications including network routing, classification, comparison and prediction[8,9].

The widely used module determination methods provide useful and clear dissections of the network, but many times do not determine the overlaps, and suffer from the 'giant component problem' failing to determine small and large modules simultaneously (Fig. S1, Table S1 and S2)[1-7]. Keeping in mind the emerging needs for an integrative approach avoiding the above problems we have developed the concept of community landscapes and a three-step framework, called ModuLand, to construct them (Fig. 1. and Fig. S2). First, we determined the 'community heaps' of the network by three community heap construction methods, called NodeLand, LinkLand and PerturLand, which we describe in detail in Methods and in the Supporting Information (SI). Since the LinkLand method behaves as a representative 'average' of the other two methods we used this method in most cases. Here the community heap was defined as the maximal connected subgraph of the starting link, where the threshold of [the sum of the weights of the links belonging to the community heap] / [number of elements in the community heap] was not allowed to decrease during the network walk providing the growth of the subgraph. Fig. 1A shows three community heaps of the network science co-authorship network[10]. All the three starting links highlighted by the arrows belong to widely collaborating, key players of the field, having a large community heap.

The 'community landscape' is generated by summing up all the community heap values of a given link of the network, and by consequently performing this summation to each link. The resulting sums may be plotted vertically over a 2D representation of the

network resulting in a 3D visual image of the community landscape as shown on Fig. 1B. The community landscape height is a centrality-type measure, since a central link both belongs to many community heaps and has large community heap values in many of them (Fig. S3). The third step of the ModuLand method-family identifies the modules of the network by finding the 'hills and mountains' of the community landscape. This is seemingly easy (we can 'see' them on Fig. 1B), but we should not forget that the position of links in their 2D network representation of Fig. 1 already reflects the information on the density and strength of their interactions. In the two implementations we worked out in detail, the TotalHill and ProportionalHill methods (Fig. 1C and Fig. S2) we start the identification of the modules by finding the module centers, i.e. the links (or plateaus of links), which have a local maximum height on the community landscape. We have to note at this point that by finding local maxima the method automatically determines the maximal number of modules giving a much simpler solution than many other methods, where the exclusion of additional dissections often needs special criteria (see Fig. S4 and Table S2). Elements and links can be assigned to the module-centers by using a number of module membership assignment methods (Section V. of the SI; Figs. S4 and S5). The most detailed module overlap information is achievable by the TotalHill module membership assignment method, as shown on Fig. 1C, where large segments of the network belong to at least two modules.

Several widely used, efficient network modularization methods[1,3] can be interpreted as parts of the ModuLand framework either by identifying the underlying community heap construction method or by identifying the community landscape directly (Section IV.4. of the SI). New modularization methods can easily be generated by taking an existing ModuLand modularization protocol, and changing any of its community heap construction, landscape generation, hill determination, or module membership assignment methods. Additionally, former methods yielding non-overlapping modules can be upgraded to overlapping modularization methods using the ModuLand approach (Section IV.4. of the SI). Optionally, a higher level hierarchical representation of the network can also be created, where the nodes of the higher level correspond to the modules of the original network, and the links of the higher level correspond to the overlaps between the respective modules (Fig. 1D, Fig. S2 and Fig. S6). This hierarchical representation can be used recursively in several steps until the whole original network is represented by a single element allowing a fast, zoom-in type analysis of large networks (Section VII. of the SI).

Even a very simple version of the ModuLand method family, the NodeLand community heap construction method correctly identified the observed split of the gold-standard Zachary karate club network[11] while uncovering a third, previously identified module and several club-members in modular overlaps (Fig. S7). Application of the more detailed LinkLand community heap construction method to the University of South Florida word association network[12] resulted in a set of modules having a highly heterogeneous degree, module size and module overlap distribution (Fig. S8), which is in agreement with earlier data (see SI) [2,3]. Enriching the binary, yes/no module membership assignment of many previous methods, the ModuLand method-family gives a continuous scale for the association of each link and element to all modules (Fig. S7). To define the number of modules of link or element the 'effective number' of modules was introduced (see Section

2

V.6.b. of the SI), which is a threshold-less, continuous measure based on the effective size of support of a probability distribution[13]. The 'effective number' of modules was proved to be a fairly stable measure in case of increasingly erroneous network data (Fig. S13. and Section VI.2. of the SI), which shows the robustness of the ModuLand method. Moreover, the ModuLand method solved the giant component problem (Fig. S4)[4] by simultaneously identifying both large and small modules of the network without the use of any preset threshold. Additionally, the ModuLand method allowed the definition of a large set of novel, topological measures characterizing e.g. the centrality and bridgeness of network elements and links (Sections IV. and V.6. of the SI).

Extending the analysis of the gold-standard Zachary karate club network, we examined the much larger University of South Florida word association network having 10,617 elements and 63,788 links[12], which was a target of a successful modularization study yielding overlapping modules[3]. This detailed analysis took 10 minutes on a computer with a 3 GHz Intel CPU. Fig. 2 shows the modular environment of the antagonym word, "terrific" and that of the heteronym, "content". The mingling colors indicate a high overlap between the modules. Importantly, the overlap of the modules with alternative meanings of the two words is much greater in the case of "terrific" than in case of "content", which is a reasonable consequence of the fact that variations of antagonistic meanings ("terrific") are often amongst our associations, while associations between differently pronounced meanings ("content") are much more seldom. Overlap between the multiple meanings of the words "bright" and "focus" (Fig. S9) is closer to that of "terrific" than that of "content". However, in case of these latter, multiple meaning words the similarly pronounced meanings are not divided into two major sections as in case of the antagonyms or heteronyms, which is again in agreement with our common knowledge.

The modular hierarchy of the high school friendship Community-44 of the Add-Health dataset[14] was uncovered using several community heap construction methods all revealing four well-distinguishable main modules with a large amount of further sub-modules (Fig. 3A and Figs. S10-S12). Girls were less likely to form multiracial modules (Fig. 3B), and the modular overlap of boys and girls significantly differed (chi-square $p < 0.00001$; Fig. 3C). These differences are in agreement with the sociological observations indicating a larger cohesiveness of friendship circles of girls than that of boys[15,16].

To test whether the ModuLand framework can identify key network elements, we calculated the change of network integrity[17] during the disintegration of the USA Western Power Grid network[18]. Elements were removed in the decreasing order of their degree, betweenness centrality and ModuLand bridgeness (measuring the bridge-like role of the elements between the modules as defined in Section V.6.d. of the SI). Fig. 4 shows that the impact of bridgeness-based element removal on network integrity was larger than that of the degree-based attacks and was well comparable to, or better than the result of betweenness centrality-based element removal. The equal-to-better performance of bridgeness-based disintegration compared to that using betweenness centrality is surprising all the more,

since the global network integrity measure corresponds extremely well to the global betweenness centrality measure[17].

Discrimination of date- and party-hubs of protein interaction networks, i.e. proteins sequentially or simultaneously interacting with a large number of neighbors, is a rather difficult task[19-24]. We hypothesized, that among date-hubs and party-hubs of similar centrality, date-hubs may have a higher bridgeness (i.e. they are more overlapping between modules of the network). This assumption was substantiated by the inter-modular position of date-hubs[19,22] and by the similarly high efficiency of bridgeness-based and date-hub-based network disintegration (cf. Fig. 4 with Fig. 2 of ref. 19 and ref. 24). The identification of the overlapping modules of a high-confidence yeast protein-protein interaction network[20] resulted in a number of modules with well-known functions (Fig. 5A and Fig. S14). We calculated the bridgeness and centrality measures of the individual proteins, and plotted these values on Fig. 5B. The separation of date- and party-hubs represented by the line of Fig. 5B classified 84 party-hubs correctly of the total of 201, and 307 date-hubs of the total of 318. This result becomes even more convincing, if we consider that 10 out of 11 incorrectly identified date-hubs (91%) and 89 out of 117 incorrectly identified party-hubs (76%) have been potentially misclassified, if comparing them to the consensus of classifications[19-23]. In conclusion, by the help of the novel measures of the ModuLand-based analysis, we were able to discriminate between date- and party-hubs, thus predicting the dynamic behavior of network elements using only the topological information of their network.

The ModuLand method-family we introduced in this paper and in part in an earlier patent application[25] is a novel, fast and robust method, which can be tailored for the special needs of the experimenter as well as for the conditions of the network studied. The method gives a comprehensible, hierarchical representation of large, real-world networks. Several key steps and especially their combination in the ModuLand method-family are novel, since (a) such a large variety of community heap-determination methods have not been integrated in any modularization methods; (b) community landscapes and their hills have never been used to determine network modules. Previous methods using local community detection or yielding overlapping modules (Table S2)[3,6] used only one or another of the approaches presented here, and did not combine any of them to community landscapes. Previous network landscape methods utilized local elements of topology[26,27], while the ModuLand method assesses a wide range of structural information. Moreover, none of the previous authors used their landscapes for module determination. The recent work of Roswall and Bergstrom[5] published during the course of the current study[25] used the probability flow of random walks to construct a map of scientific communication. This method is similar to our PerturLand community heap construction method, but its application in ref. 5 yields non-overlapping modules.

The extensive and rich overlaps, network hierarchy, as well as the novel centrality and bridgeness measures uncovered by the ModuLand method can be used for the identification of long-range, stabilizing weak links, for the determination of the recently described creative, trend-setting elements governing network development and evolution[28], for prediction of missing links or elements, for network classification and for

the design of efficient information transfer to name only a few of the many possibilities. Module overlaps might play a key role in the disconnection and synchronization of modules of complex systems, and their re-assembly during and after crisis, respectively. We invite our colleagues to design novel versions of the framework we gave, and to explore the above and other examples.

Methods

Networks

Network science co-authorship network. The giant component of the undirected, un-weighted network science co-authorship network contained 379 elements and 914 links[10]. *Karate club social network.* The weighted and undirected social network of a karate club has been reported by W. Zachary[11] containing 34 elements and 78 links. As the members of the karate club have split into two factions later, the network became a gold-standard of module determination methods[1-3,5-7]. *Word association network.* The giant component of Appendix A of the University of South Florida word association network (http://www.usf.edu/FreeAssociation/)[12] with removed link directions contained 10,167 elements and 63,788 weighted links, where weight refers to the association strength (see Section I.3. of the SI). *School friendship network.* The giant component of the high school friendship Community-44 of the Add-Health database (http://www.cpc.unc.edu/projects/addhealth)[14] with removed link directions contained 1,127 elements and 5,096 weighted links, where weights represent the strengths of friendships (see Section I.4. of the SI). *Power-grid network.* The unweighted and undirected network of the USA Western Power Grid[18] contained 4,941 elements and 6,594 links. *Yeast protein-protein interaction network.* The giant component of the un-weighted and undirected yeast protein-protein interaction network[20] contained 2,444 elements and 6,271 links, covering approximately half of the yeast genome and the most reliable ('strongest') ~3% of the expected number of total links.

Brief description of the 'NodeLand', 'LinkLand' and 'Perturland' community heap construction methods

The ModuLand module determination protocol gives an integrative approach, where modifications of the various parameters used in the three major steps (community heap formation, community landscape construction and module membership assignment, the latter including the hill determination step) may result in a large number of special implementations. We have developed three community heap construction methods (NodeLand, LinkLand and PerturLand), one trivial landscape construction method summing the community heap values over the links of the network, and two module membership assignment methods (ProportionalHill and TotalHill, each utilizing the PeakHill hill determination method). We describe these methods in the SI in detail, and list only the major steps of the three community heap determination methods here. As discussed in Section IV. of the SI, the best community heap determination method depends on the input of network data. From the three community heap determination methods the perturbation propagation-based PerturLand method gives the most detailed and robust results. However, the LinkLand and the more basic NodeLand methods exploring the densely connected region of the starting link or element, respectively, are faster, require no parameter tuning, and give us accurate enough results with even inaccurate link weights. A detailed description of these methods is given in Sections IV. and V. of the SI. The Linux-based computer programs of the ModuLand-related methods (7 Mbytes) and a Windows-based application (240 Mbytes) can be downloaded from here: www.linkgroup.hu/modules.php.

References
1. Girvan, M. & Newman, M. E. J. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA* **99,** 7821–7826 (2002).
2. Radicchi, F., Castellano, C., Cecconi, F., Loreto, V. & Parisi, D. Defining and identifying communities in networks. *Proc. Natl. Acad. Sci. USA* **101,** 2658–2663 (2004).
3. Palla, G., Derenyi, I., Farkas, T. & Vicsek, T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **435,** 814–818 (2005).
4. Fortunato, S. & Barthélemy, M. Resolution limit in community detection. *Proc. Natl. Acad. Sci. USA* **104,** 36–41 (2007).
5. Rosvall, M. & Bergstom, C. T. Maps of random walks reveal community structure in complex networks. *Proc. Natl. Acad. Sci. USA* **105,** 1118–1123 (2008).
6. Nepusz, T., Petróczi, A., Négyessy, L. & Bazsó, F. Fuzzy communities and the concept of bridgeness in complex networks. *Phys. Rev. E* **77,** 016107 (2008).
7. Fortunato, S. Community detection in graphs. http://arxiv.org/abs/0906.0612 (2009).
8. Colizza, V., Flammini, A., Serrano, M. A. & Vespignani, A. Detecting rich-club ordering in complex networks. *Nature Physics* **2,** 110–115 (2006).
9. Guimerá, R., Sales-Pardo, M. & Amaral, L. A. N. Classes of complex networks defined by role-to-role connectivity profiles. *Nature Physics* **3,** 63–69 (2007).
10. Newman, M. E. J. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E* **74,** 036104 (2006).
11. Zachary, W. W. An information flow model for conflict and fission in small groups. *J. Anthropol. Res.* **33,** 452–473 (1977).
12. Nelson, D. L., McEvoy, C. L., Walling, J. W. & Wheeler, J. W. The University of South Florida homograph norms. *Behav. Res. Meth. Instrum.* **12,** 16–37 (1980).
13. Grendar, M. Entropy and effective support size. *Entropy* **8,** 169-174 (2006).
14. Moody, J. Race, school integration and friendship segregation in America. *Am. J. Sociol.* **107,** 679–716 (2001).
15. Clark, M. L. & Ayers, M. Friendship similarity during early adolescence: gender and racial patterns. *J. Phsychol.* **126,** 393–405 (1992).
16. Baerveldt, C., van Duijn, M. A. J., Vermeij, L. & van Hemert, D. A. Ethnic boundaries and personal choice: Assessing the influence of individual inclinations to choose intra-ethnic relationships on pupils' networks. *Soc. Networks* **26,** 55–74 (2004).
17. Latora, V. & Marchiori, M. Efficient behavior of small-world networks. *Phys. Rev. Lett.* **87,** 198701 (2001).
18. Watts, D. J. & Strogatz, S. H. Collective dynamics of ' small-world' networks. *Nature* **393,** 440–442 (1998).
19. Han, J. D. *et al.* Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* **430,** 88–93 (2004).
20. Ekman, D., Light, S., Bjorklund, A. K. & Elofsson, A. What properties characterize the hub proteins of the protein-protein interaction network of *Saccharomyces cerevisiae*? *Genome Biol.* **7,** R45 (2006).
21. Kim, P. M., Lu, L. J., Xia, Y. & Gerstein, M. B. Relating three-dimensional structures to protein networks provides evolutionary insights. *Science* **314,** 1938–1941 (2006).
22. Komurov, K. & White, M. Revealing static and dynamic modular architecture of the eukaryotic protein interaction network. *Mol. Systems Biol.* **3,** 110 (2007).

23. Bertin, N. *et al.* Confirmation of organized modularity in the yeast interactome. *PLoS Biol.* **5,** e154 (2007).
24. Yu, H., Kim, P. M., Sprecher, E., Trifonov, V. & Gerstein, M. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput. Biol.* **3,** e59 (2007).
25. Kovacs, I., Csermely, P., Korcsmaros, T. & Szalay, M. Method for analyzing the fine structure of networks. Patent application WO 2007/093960 (2006).
26. Eckmann, J. P. & Moses, E. Curvature of co-links uncovers hidden thematic layers in the World Wide Web. *Proc. Natl. Acad. Sci. USA* **99,** 5825–5829 (2002).
27. Axelsen, J. B., Bernhardsson, S., Rosvall, M., Sneppen, K. & Trusina, A. Degree landscapes in scale-free networks. *Phys. Rev. E.* **74,** 036119 (2006).
28. Csermely, P. Creative elements: network-based predictions of active centres in proteins, cellular and social networks. *Trends Biochem. Sci.* **33,** 569–576 (2008).
29. Batagelj, V. & Mrvar, A. Pajek: A program for large networks analysis. *Connections* **21,** 47–57 (1998).
30. Gansner, E. R. & North, S. C. An open graph visualization system and its applications to software engineering. *Softw. Pract. Exp.* **30,** 1203–1233 (1999).
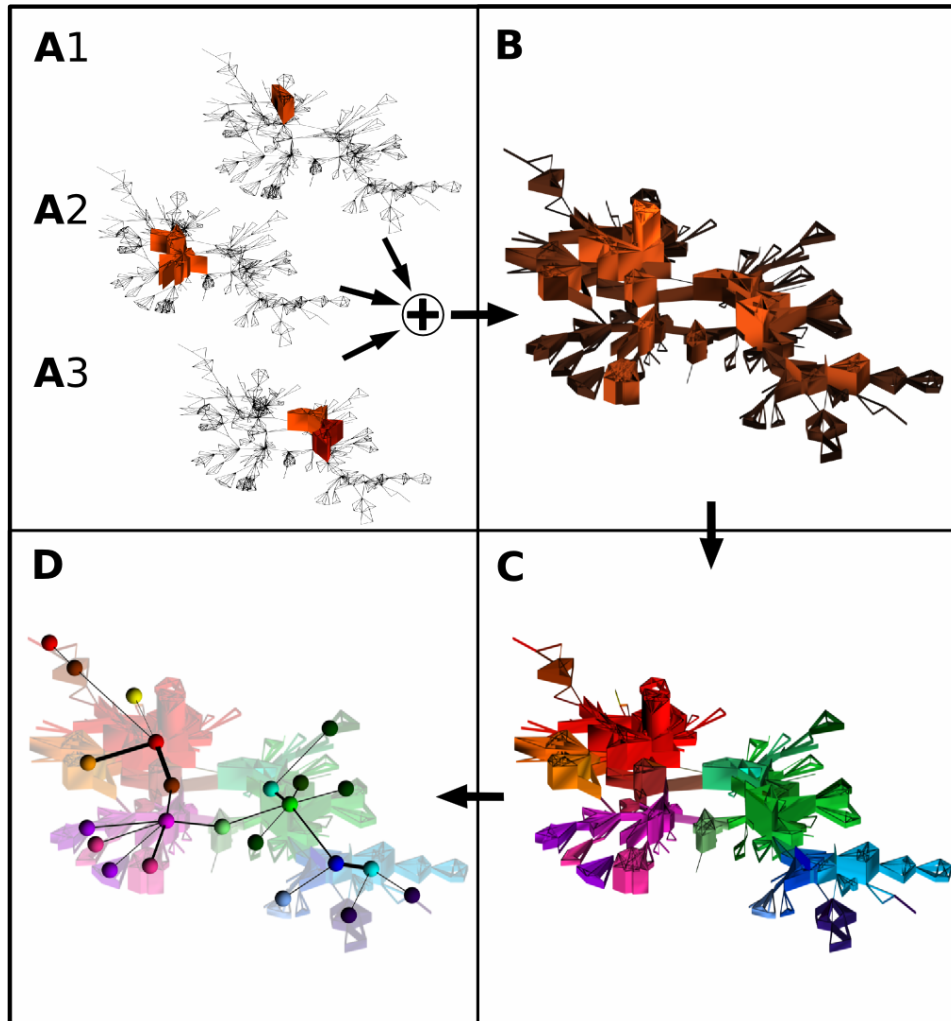
Figure 1 Description of the ModuLand method-family. *A1-A3*, First, the community heap of each link (or element) of the network are identified. If a link is in the 'middle' of a module, it will be a part of many community heaps (all the three widely collaborating, highlighted author-pairs, shown by the arrows are from this category). On the contrary, links at module 'edges' will belong to a few community heaps only. *B,* Next, the community landscape is constructed by summing up the community heap values. The hills of the community landscape correspond to the modules of the network. *C*, Last, modular centers are identified as the links at the maxima of the community landscapes, and memberships of links in all network modules are determined. *D*, Optionally, a higher level hierarchical representation of the network can be created, where elements of the higher level correspond to modules of the original network, and links of the higher level correspond to overlaps between the respective modules. On the vertical axes community heap values (panel A), or community landscape values (panels B, C and D) of the links are shown. For this illustrative example we used the network science co-authorship network[10] without link weights using the LinkLand and TotalHill methods. The network was visualized using the Pajek program[28] with the Kamada-Kawai algorithm supplemented with a custom Blender script. Community heaps of panels *A1* or *A2* belong to the Barabási—Vicsek or Girvan—Newman author-pairs, respectively. Panel *A3* shows the merged community heap of the Arenas—Pastor-Satorras and Guimera—Amaral co-authorship links. Links and nodes of panels *C* and *D* are colored in proportion of the colors of the modules they belong.
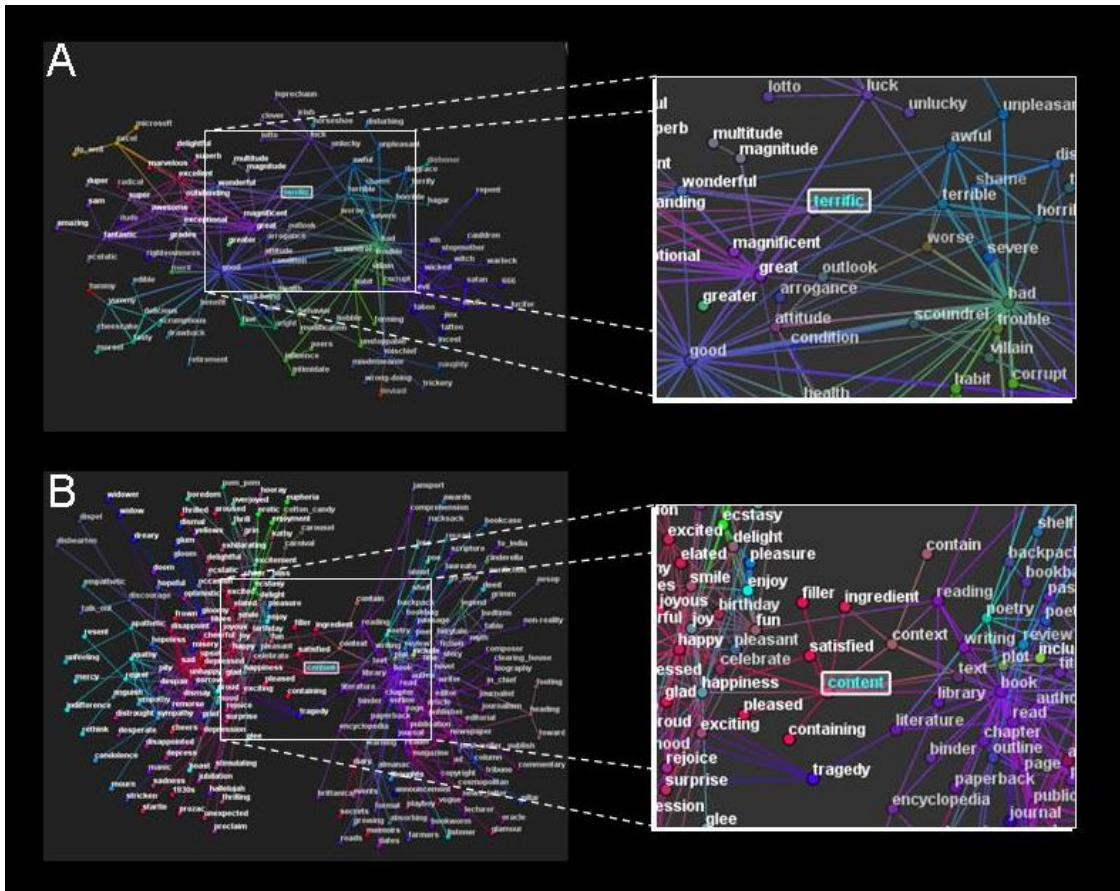
Figure 2 Overlapping modules of a word-association network. Modules of the University of South Florida word association network[12] were determined using the LinkLand community heap construction method and the TotalHill module membership assignment method. The network was laid out using the Kamada-Kawai algorithm of Graphviz[30] and visualized using a custom program written in Python language using OpenGL graphics. Links were colored in proportion to the colors of the modules they belong. *A*, Modules around the antagonym word, "terrific". *B*, Modules around the heteronym word, "content". In addition to the selected words "terrific" and "content" similar words above a similarity threshold of 10% are also shown with a contrast corresponding to their degree of similarity. The extent of similarity between two words was calculated as the sum of the two pair-wise minima of their unity-normalized module membership vector giving the membership assignment strength of the given word to all modules of the network (for more details see Section V.6.e. of the SI).
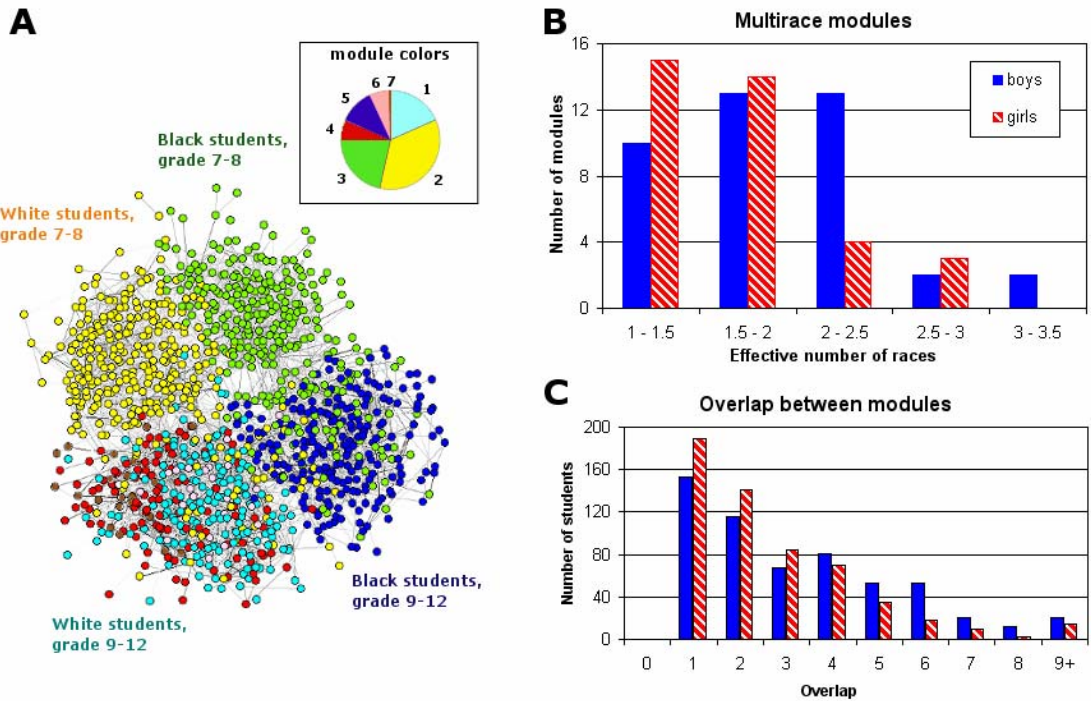
Figure 3 Overlapping modules of a school-friendship network. We have determined the modular structure of Community-44 of the Add Health survey[14] using the LinkLand community heap determination method together with the ProportionalHill module membership assignment method. *A,* Modules of Community-44. The school friendship network was visualized using the Pajek-program[28] with the Kamada-Kawai algorithm. Elements represent the individual students, and were colored in proportion to the colors of the different friendship modules they belong. We show the modular structure of the second hierarchical level having 7 modules. The *inset* of Panel A shows color-codes of the modules with an area proportional to the size of the respective module. *B,* The number of network modules in case of boys (blue, solid bars) and girls (red-white hatched bars) with mixed racial contents in the first hierarchical level. The extent of mixed racial content was monitored using the 'effective number of races' (Section V.6.b. of the SI) with a bin-size of 0.5. *C,* The number of boys (blue, solid bars) and girls (red-white hatched bars) having different overlaps in friendship circles as determined in the first hierarchical level with a bin-size of 1. Overlap was measured as the 'effective number' (Section V.6.b. of the SI) of modules of the given student.
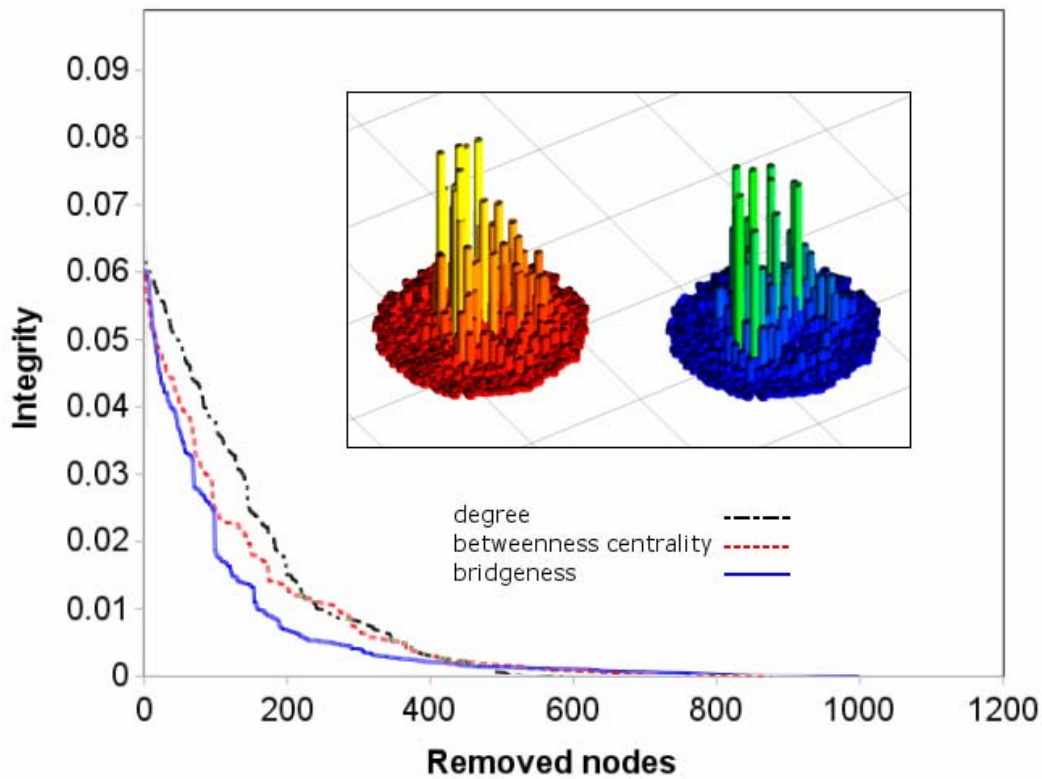
Figure 4 Determination of key elements of the USA Western Power Grid network. The figure shows the decreasing integrity of the USA Western Power Grid network[18] as a function of the number of elements removed. Elements were removed in the order of their decreasing degree (black alternating dashes and dots) betweenness centrality (red dashed lines)[1] or 'bridgeness' (solid blue lines), where 'bridgeness' measures the overlap of the given element between different modules as described in detail in Section V.6.d. of the SI. Network integrity has been calculated after Latora and Marchiori[17]. Bridgeness was calculated from the modular structure of the lowest hierarchical level as determined by the LinkLand community heap construction method and the TotalHill module membership assignment method. On the vertical axis of the *insets* the betweenness centrality (left, color-coded from red to yellow) and bridgeness (right, color-coded from blue to green) of the elements of the USA Western Power Grid network are shown. Networks on the *insets* were laid out using the Pajek-program[29] with the Kamada-Kawai algorithm and visualized with a custom Blender script.
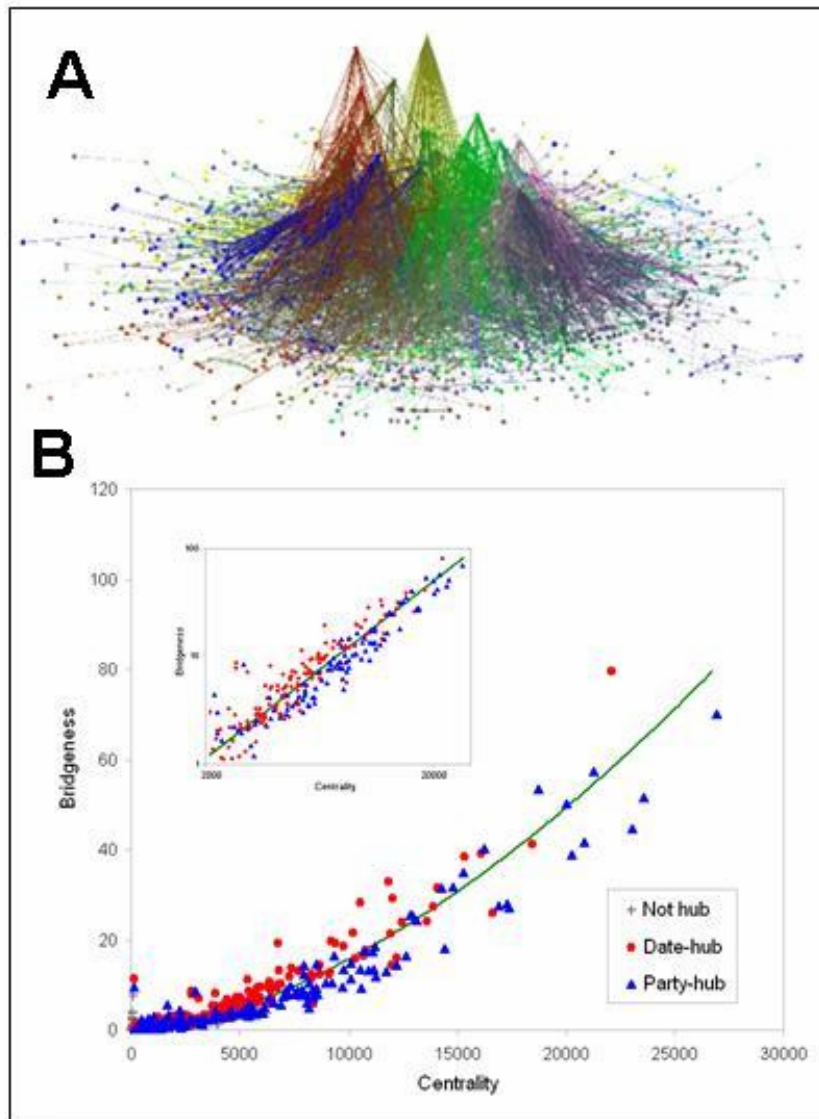
**Figure 5 Prediction of the dynamical behavior of network elements: segregation of date-
and party-hubs based on their modular overlaps.** Overlapping modules of the yeast protein-
protein interaction network of Ekman et al.[20] were identified using the LinkLand community heap
determination method with the TotalHill module membership assignment method using the
modular structure of the lowest level of hierarchy. *A*, A 3D view of the yeast protein-protein
interaction network. The underlying 2D network layout was set by the Pajek-program[29] with the
Kamada-Kawai algorithm. The vertical positions reflect the community landscape values of the
elements on a linear scale. Elements were colored as the module of their maximum membership.
B, Centrality and bridgeness of yeast date- and party-hubs. Hubs having more than 8 neighbors
and non-hubs with less neighbors were positioned on the scattergram according to their
ModuLand centrality (x-axis, the height of the community landscape) and ModuLand bridgeness
(y-axis) as defined in Section V.6.d. of the SI. Date- and party-hubs are marked with red circles and
blue triangles, respectively, while non-hub proteins are represented by gray crosses. The *inset*
shows a double logarithmic plot of hubs with large centrality.