Visual Statistics for Collections of Clustered Graphs

Ulrik Brandes
Dept. Computer & Information Science
University of Konstanz
Ulrik.Brandes@uni-konstanz.de

Jürgen Lerner*
Dept. Computer & Information Science
University of Konstanz
lerner@inf.uni-konstanz.de

Miranda J. Lubbers
Dept. Social & Cultural Anthropology
Autonomous University of Barcelona
mirandajessica.lubbers@uab.es

Chris McCarty
Bureau of Economic and Business Research
University of Florida
chrism@bebr.ufl.edu

José Luis Molina
Dept. Social & Cultural Anthropology
Autonomous University of Barcelona
joseluis.molina@uab.es

ABSTRACT

We propose a method to visually summarize collections of networks on which a clustering of the vertices is given. Our method allows for efficient comparison of individual networks, as well as for visualizing the average composition and structure of a set of networks. As a concrete application we analyze a set of several hundred personal networks of migrants. On the individual level the network images provide visual hints for assessing the mode of acculturation of the respondent. On the population level they show how cultural integration varies with specific characteristics of the migrants such as country of origin, years of residence, or skin color.

Keywords: Clustered graph visualization, social network analysis, acculturation.

1 INTRODUCTION

Social networks encode important information about the social contacts of individuals and the overall structure of the community. A powerful way to explore and represent such networks is the visualization of graphs. Well-designed network images reveal important structural features and provide means to visually compare two or more networks, thereby revealing differences between communities or individual actors. However, graph visualization reaches certain limits if the networks become very large or if a huge number of networks has to be displayed simultaneously, since the drawings inevitably become too crowded.

To overcome these limitations, a growing number of techniques to draw *clustered* graphs has been proposed. These methods visualize a potentially large network by exploiting a given clustering of the vertices and grouping vertices in the same cluster visually together. Besides reducing the visual complexity, clustered graph visualizations are often better interpretable than images obtained without any clustering: For instance, an image that shows how classes of actors with particular attribute values (such as gender, age, skin-color, or religion) are connected indicates how actor characteristics influence the creation of ties in the analyzed community. Thus, the analyst obtains information that might generalize beyond the given network sample. However, existing methods for clustered graph visualization are inconvenient for (and, in fact, they are not targeted at) the task of drawing *collections* of clustered graphs to allow for easy comparison across networks and for visual averaging over a set of networks.

In this paper we propose a method for drawing clustered graphs having two application scenarios in mind. First, to draw dozens or hundreds of clustered graphs simultaneously so that we can easily recognize differences and similarities in network composition and structure. Second, to draw an aggregated view over a potentially large number of networks, showing the trend (statistical average) and dispersion (statistical variability) in the analyzed population. Our images abstract from individual vertices and show only class sizes, average connectivity within and between classes, and (if desired) mean values and variances of these indicators. This decision is crucial to obtain small but readable visual summaries of networks and to make simple comparison between disjoint networks possible at all. Furthermore, our method is very efficient in terms of running time (linear in the total number of vertices and edges) and can therefore be applied to very large networks or large numbers of networks. Last but not least, our visualization technique is conceptually very simple so that it can be expected to introduce only little artefacts and is also usable by practitioners that are not experts in graph drawing algorithms.

As a concrete application we present a visual analysis of personal networks of several hundred migrants to the USA and to Spain. On the individual level these networks provide a visual measure for *acculturation* of migrants to the host culture—improving traditional measures of acculturation that do not take into account network structure. On the aggregated level, our images show the average network composition and structure in purposefully chosen sub-samples, thereby revealing how the mode of acculturation depends on the country of origin, time of residence, and skin color of the respondents.

1.1 Related Work

The visualization of graphs (or networks) is an important part of data and information visualization and a large number of methods to draw graphs has been developed (see e. g., [6, 11]). Although efficient methods are presently able to layout (i. e., to compute coordinates of vertices and edges) graphs that have several hundreds of thousands of vertices, the drawings of such large graphs are too crowded to be captured by the human cognitive system. To obtain simpler images that still reveal essential information about the network, several researchers propose to make use of a given clustering of the graph. Proposals include visualization techniques for hierarchically clustered graphs [7, 10], dynamic drawing of clustered graphs [9], systems for visual navigation trough clustered graphs [8, 15], visual interpolation between different degrees of clustering or levels of detail [1, 13], and algorithms to obtain clustered graph layouts that optimize certain esthetic criteria [2].

Out of the abovementioned papers only the method from Frishman and Tal [9] is designed to draw *several* clustered graphs such that changes from one image to the next can be easily seen. Note that the method from [9] works only for graphs with largely overlapping vertex sets—a task that falls into the area of *simultaneous graph drawing* (see, e. g., [12] and references therein). In contrast,

*corresponding author

our method is aimed at visualizing collections of clustered graphs whose vertex sets may be disjoint and where only a one-to-one correspondence between the cluster labels is given.

Further previous work, which is related to acculturation and network analysis, is overviewed in Sect. 2.

Outline of paper. We introduce an exemplary application in Sect. 2. Networks of individuals are drawn by the methods described in Sect. 3, whereas Sect. 4 details how the average and dispersion of a set of networks is defined and visualized. In Sect. 5, we illustrate how average network images give deep insight into a collection of personal networks. We close with a discussion of results and future work.

2 Example Application: Towards a Network Measure for Acculturation

Although this paper presents a method for the visualization of networks and is not intended to provide conclusive results for the social sciences, we nevertheless dedicate this section to introduce a real-world application. By describing how our visualization technique helps in the task of analyzing personal networks of migrants, we provide hints how it can be applied in other scenarios. Besides, we claim that the study of acculturation is interesting and relevant on its own.

Acculturation refers to phenomena which result if different cultures come into contact [14]. Recently, the term acculturation is often used to denote the integration of migrants into a host culture. In this paper, we understand acculturation in this second, more restrictive usage. Migration probably took place throughout the history of mankind but, due to advances in transportation and communication means, it is increasing in quantity and speed in recent decades, explaining an increased interest in understanding and measuring acculturation.

Acculturation scales are not one-dimensional measures ranging from (say) not integrated to fully integrated, but make finer distinctions. Berry [3] defined four strategies (modes) of acculturation based on two dimensions for cultural affinity, see Fig. 1.

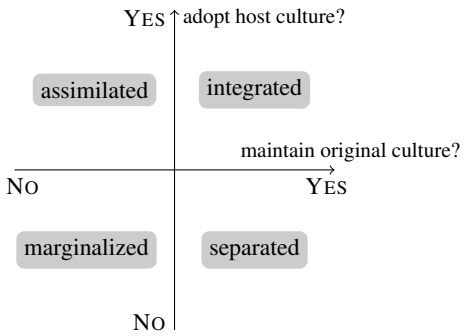


Figure 1: Four traditional acculturation strategies (shaded) defined by the migrant's affinity to her original culture and to the host culture [3].

The two dimensions for cultural affinity are traditionally measured by responses to pairs of statements such as the following (intended to measure the integration of Mexican migrants to the USA; small sample taken from [5]).

1. I speak English / I speak Spanish
2. I enjoy English language TV / I enjoy Spanish language TV
3. My friends now are of Anglo origin / ... Mexican origin

A criticism to these kind of measurements for acculturation is that they are too much focused on the respondent's *individual attributes* (such as languages spoken or media consumed) and do not take sufficiently into account the migrant's social network. While some statements (e. g., the third in the list above) are concerned with *network composition* (who is in the network), the *network structure* (how are they connected) is not taken into account at all. Suppose that a migrant knows both, many Mexicans and many Anglo Americans. Still it makes a big difference whether these two groups are well connected to each other (one homogeneous bicultural network), or whether they are separated (the migrant lives in two culturally different societies). The long-term objective of this project[1] is to measure and understand acculturation by simultaneously taking into account individual attributes, network composition, and network structure. Although the use of personal networks is not so established in the study of acculturation, many results in social network analysis (see [4, 17] for an overview) demonstrate the importance of network structure.

Clearly, when incorporating the migrants' personal networks into studies of acculturation (and not only their attributes), the data sets describing individuals become much more complex. For this reason, good visual support for exploring the empirical network data becomes even more important.

2.1 Empirical Data Set

For the present paper we used a data set obtained by interviewing 500 migrants (alternatively referred to as *respondents* or *egos*) to the USA and to Spain, originally coming from different South-American, Middle-American, and African countries. Each respondent was asked to provide the following four types of information with the help of the EgoNet[2] software.

1. **(questions about ego)** 70 questions about the respondent herself, including age, skin color, years of residence, questions from traditional acculturation scales (such as [5]), and health related questions.
2. **(name generator)** A list of 45 persons (referred to as *alters*), personally known to the respondent. The alters are the vertices in the respondent's personal network.
3. **(questions about alters)** 12 questions about each of the 45 alters, including country of origin , country of residence, skin color, and type of relation to ego.
4. **(ties between alters)** For each of the 990 undirected pairs of alters, the evaluation whether they know each other. The three possible choices were "very likely," "maybe," or "unlikely" and we introduced an edge in the network only if the respondent chose "very likely."

Note that, although each network is seemingly rather small (only 45 vertices), the number of networks (500) and the many ego-attributes and alter-attributes make up a quite large and semantically rich data set.

3 Visual Comparison of Individual Networks

Suppose that we are given a set of N undirected graphs $G_1 = (V_1, E_1), \ldots, G_N = (V_N, E_N)$ and a rule to partition the vertex sets $V_1, \ldots, V_N$. Our goal in this section is to draw the N clustered graphs simultaneously so that we can easily compare them visually, find differences and similarities among them, and detect networks having specific characteristics. In our drawings we do not show individual vertices (the elements of the V_i) but only the classes and the average connectivity between and within them. This decision is due to two reasons. First, by doing so it is possible to obtain informative images of many networks on small space (compare Fig. 2). Second,

[1] see `http://www.egoredes.net` for a description of the project
[2] see `http://www.mdlogix.com/egonet.htm`

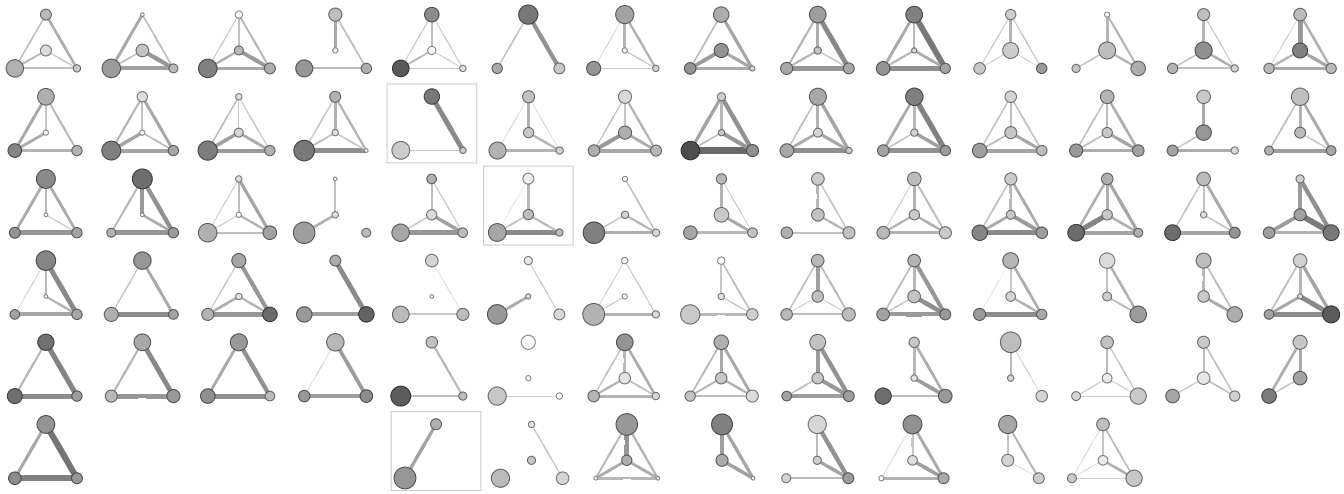


Figure 2: Personal networks of 79 Argentinian migrants to Spain. The four nodes of each network correspond to the four classes defined in Sect. 3.1 (also see Fig. 3). Node size reflects class size, darkness of a node reflects average connectivity within the class, width and darkness of a tie between nodes reflects average connectivity between the two classes (compare Sect. 3.2 and Fig. 4).

in the application at hand the vertex sets are disjoint and the individual alters are normally unknown to the analyst (so that comparison on the individual level is hard to achieve and hard to interpret), but the classes have a well-defined interpretation (see Sect. 3.1) that generalizes across different networks.

Visualizing the class-level networks is done in two steps: defining the actor classes (Sect. 3.1) and defining how strongly two classes are connected (Sect. 3.2).

3.1 Definition of Classes

The partitioning of the N vertex sets $V_1, \ldots, V_N$, each into k classes $V_i = C_1(i) \cup \ldots \cup C_k(i)$ (some of which may be empty), is required to be consistent in the sense that for two different networks G_i and G_j and an index (a class label) p the class $C_p(i)$ corresponds to class $C_p(j)$. In our application, this labeling is achieved by specific attribute values as it is explained below. (Note that the requirement of consistently labeled vertex classes is not only satisfiable in socal network analysis. For instance, classes of Web-pages might be labeled by domain names, classes of Wikipedia pages by membership in certain categories, etc.) The layout (coordinates) of the k classes is required to be the same for each network. This decision is crucial for the applications that we envision here, since it allows for quick and simple comparison without forcing the analyst to read class labels. The layout of classes can either be chosen by the analyst who knows about the semantics of the different classes (as in our case), or it can be computed (e. g., by force-directed graph layout techniques) on an aggregated view of the whole set of networks, so that classes that are often well-connected are drawn close together.

In our specific application we take (in almost all cases, except the bottom row in Fig. 11) the definition of actor classes derived in the following. Since we want to measure the migrant's affinity to both her original culture and the host culture, the most important distinction is between actors originating from country of origin (compatriots) and actors from the host country. However, while exploring the data it turned out that the class of compatriots should be further refined. Typically, one part of these actors still lives in the country of origin and others live in the host country. (As a matter of fact, migrants often become more easily acquainted to compatriots abroad—even if they did not know them while still being in their home country.) This distinction is important because (normally) the migrant interacts with the former class only via the distance (telephone, email, etc.), whereas people currently living in the same host country make up the real social community of the migrant and play a role in her everyday life. Finally, a default class contains all actors that are neither born in the country of origin, nor in the host country.

The concrete definition, labeling, and layout of these classes is given in Fig. 3, where we assume for simplicity that the respondent is an Argentinian that migrated to Spain. Note that for a (say) Puerto Rican that migrated to the USA, the definition of classes changes in an obvious way, although their labels (ORIGIN, FELLOWS, HOST, and TRANSNATIONALS) stay the same. We show examples of networks that have been partitioned into more classes in Fig. 11.

(origin) born in Argentina, not living in Spain

(fellows) born in Argentina, living in Spain

(host) born in Spain

(transnationals) all others

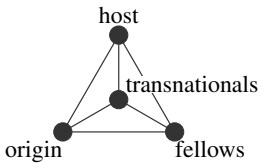


Figure 3: *Left*: Definition and labeling of specific actor classes, assuming that the respondent migrated from Argentina to Spain. *Right*: A fixed layout (coordinates) for these classes.

The three corners of the outer triangle in Fig. 3 give a summary of the network composition: If the class ORIGIN is very large, the migrant is not only focused on her original culture, she also interacts mostly with people that are not living in her country of residence. If the network is mostly composed of FELLOWS, the migrant's social network is located where she is currently living but she is still attached to her original culture. A migrant whose network is dominated by the HOST class appears to be integrated or even assimilated (compare Fig. 1). The class of TRANSNATIONALS has deliberately be put on the (neutral) position in the middle, as their cultural orientation is not obvious. Furthermore, migrants whose networks are dominated by transnationals are rare in our data set. Respondents having uncommonly many/few alters in specific classes (as well as more balanced networks) can easily be identified, e. g., in Fig. 2.

3.2 Intra-class and Inter-class Ties

Besides the relative class-sizes (network composition), it is important to know how actors in various classes are connected to each other (network structure). In the following we derive an indicator for how strongly actors in a class $A \subseteq V$ are on average connected to actors in a class $B \subseteq V$ in the network $G = (V,E)$. (Class A and B may be identical.)

A basic un-normalized measure would be to count the number of links connecting A and B:

$$e(A,B) = |\{(a,b) \in E\,;\, a \in A \text{ and } b \in B\}| \quad . \tag{1}$$

However, when using this measure larger classes would (normally) be stronger connected, so that it has to be normalized appropriately.

A commonly used normalized measure is the *density* of ties between A and B, i. e., the fraction of all realized ties over all possible ties: $e(A,B)/(|A| \cdot |B|)$. Unfortunately, the density normalizes too much so that two very small classes (that are connected by only a few ties) reach unjustified high values. For instance, if the sizes of classes A and B are equal to two and A and B are connected by only two ties, the density is 0.5 (and thus quite high), although an A-vertex has on average only one B-neighbor. On the other hand, if the sizes of classes A and B are equal to 20, then an A-vertex needs on average 10 B-neighbors to make the density equal to 0.5. Furthermore, note that for sparse graphs the density tends necessarily to zero when the class-sizes increase.

Taking into account the considerations above, a reasonable measure of adjacency between classes (which is already close to our final definition) seems to be the number of B-neighbors that an A-vertex has on average, i. e., the value $e(A,B)/|A|$. However, this measure has the counterintuitive property that the network on the class-level may be asymmetric (the weight of an edge and its reverse may differ), even if the original graph is symmetric (undirected).

A measure that does not have this disadvantage (but that is still built on the idea of average number of neighbors between classes) is obtained by taking the geometric mean of the class sizes of A and B instead of $|A|$ in the denominator, resulting in our final definition of the *weight* how strongly two classes A and B are connected:

$$\omega(A,B) = \frac{e(A,B)}{\sqrt{|A| \cdot |B|}} \quad . \tag{2}$$

Even if in some applications other definitions for the edge weights between classes might be appropriate, we adhere in the whole paper to that given in (2).

3.2.1 Visual Representation

The class sizes and weights of intra-class ties and inter-class ties are visually encodes as described in the following, see Fig. 4 for illustration. A class C is drawn as a circle whose area size is proportional to $|C|$ (the size of C) and hence whose radius is proportional to the square-root of $|C|$. The circles (nodes) are filled with a grey color whose darkness is proportional to the weight of the corresponding intra-class tie. The nodes are connected by edges whose width is proportional to the weight of the corresponding inter-class tie (the color of these edges is chosen by the same rule as for the intra-class ties). The specific representation of tie weights has been chosen because it goes smoothly with other visualization techniques for clustered graphs (such as [1, 2]): if vertices belonging to the same class are connected by many edges, then the area of this class appears to be intensively colored with the edge color. Similarly, if many edges connect vertices from two different specific classes, then the corresponding edge bundle gets wider and/or denser.

Figure 2 shows that we can easily identify instances with specific characteristics in drawings of dozens of networks. Not only do sizes

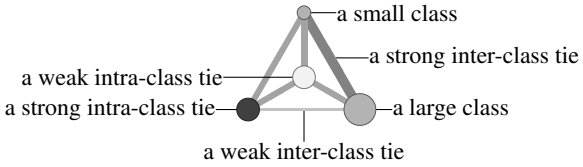


Figure 4: Illustration for the visualization of class size and tie weight.

of specific classes (network composition) vary enormously between some instances, we can also detect large differences in the relative tie-weights (network structure). For instance, the three networks in Fig. 5 show large differences with respect to which communities are connected to each other. In the lefthand side the FELLOWS (albeit small in number) are well-connected to the HOST-class, whereas the large ORIGIN-class is quite separated from the rest. Thus this migrant's network decomposes with respect to where the actors are currently living. In contrast, the strongest inter-class tie in the network in the middle of Fig. 5 is between ORIGIN and FELLOWS and, thus, this network decomposes with respect to where the actors are originally from. Finally, in the network in the righthand size of Fig. 5 only the two classes ORIGIN and HOST are non-zero and, in addition, they are well-connected. This network structure is quite rare in our sample, since typically the tie between these two classes is rather weak. In fact, these two groups neither have the same roots, nor are they living in the same country.

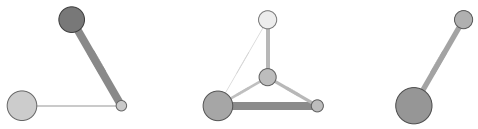


Figure 5: Three selected instances from Fig. 2 that show large differences in the adjacency structure between classes.

4 Tendency and Dispersion in a Population

The images developed in Sect. 3 are convenient to compare personal networks of individuals. However, research in the social sciences is often targeted at learning about trends in societies. For instance, a research question could be: "are migrants from South-American countries *on average* better integrated in Spain than migrants from African countries?" A related kind of question is: "do individuals from a particular country behave rather similar (low variability) or can we find large differences among them (high variability)?" This section is about visualizing tendency (statistical average) and dispersion (statistical variability) in a population. The definition of average and variability of a set of clustered networks will be based on established statistical notions such as arithmetic mean, standard deviation, median, and upper and lower quartiles. Nevertheless some care has to be taken to define these measures for the average tie weights.

4.1 Definition of Average and Variability

4.1.1 Arithmetic Mean of Clustered Networks.

If X is a random variable for which we have N observations $x_1, \ldots, x_N$, then the arithmetic mean of the N observations is the real number $\mu(x) = \sum_{i=1}^{N} x_i/N$. Similarly, if we have a sample of N clustered graphs, we want to define the arithmetic mean to be a clustered graph that is $1/N$ times the sum over the sampled graphs, compare Fig. 6.

All we have to do to make this idea precise is to define the sum of two clustered networks (an thereby the sum of an arbitrary num-

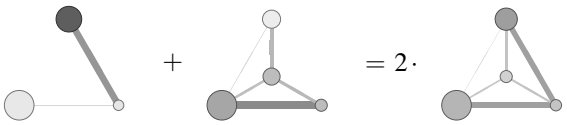


Figure 6: Arithmetic mean (*right*) of the two networks on the lefthand side of the equation. Note that both, the class sizes as well as the tie weights are averages of the corresponding values in the summands. The image on the righthand side does not visualize the variability (compare Sect. 4.2).

ber of networks). Furthermore, considering that the class-level networks are just arrays of numbers, this seems to be quite simple. Actually, it is straightforward for the class sizes and only slightly more complicated for the tie weights. So let $G_1 = (V_1, E_1), \ldots, G_N = (V_N, E_N)$ be a set of graphs whose vertex sets are all partitioned into k classes $V_i = C_1(i) \cup \ldots \cup C_k(i)$ and let p be a class-label. The size of class p in the arithmetic mean of the N graphs is defined to be

$$\mu|C_p| = \sum_{i=1}^{N} |C_p(i)|/N \ .$$

Now let $p,q = 1,\ldots,k$ be two indices of classes. The tie-weight between the p'th and the q'th class in the arithmetic mean is *not* defined to be the mean of the individual weights, $\sum_{i=1}^{N} \omega(C_p(i), C_q(i))/N$, as these weights are averages themselves (compare (2)). Before explaining how the mean tie weight is actually defined, we first illustrate why the last-mentioned formula would be a bad choice and how a better measure can be derived. Suppose we want to average over two networks G_1 and G_2, where the classes $C_p(1)$ and $C_q(1)$ have both size 10 and are connected by 100 edges, while the classes $C_p(2)$ and $C_q(2)$ have both size one and are connected by no edge. The naïve formula $(\omega(C_p(1), C_q(1)) + \omega(C_p(2), C_q(2))/2$ would determine an average weight of (only) five for the tie (C_p, C_q), although 20 out of 22 vertices are connected each by ten edges to the other class. Thus, the two vertices from network number two would get a disproportionally high influence on the average. On the other hand, if we consider the disjoint union $G_{1\cup 2} = G_1 \cup G_2$, then the classes $C_p(1\cup 2)$ and $C_q(1\cup 2)$ have both size 11, are connected altogether by 100 edges, and thus the weight of the tie $(C_p(1\cup 2), C_q(1\cup 2))$ is $100/11$, which is actually the average number of q-neighbors of a p-vertex.

Our measure for the mean tie weight is built on this idea of considering the disjoint union of the N networks. More precisely, we compute first the mean of the un-normalized edge counts (1)

$$\mu e(C_p, C_q) = \sum_{i=1}^{N} e(C_p(i), C_q(i))/N$$

and normalize by the geometric mean of the average class sizes to obtain the mean tie weight, $\mu\omega(C_p, C_q)$, i. e.,

$$\mu\omega(C_p, C_q) = \frac{\mu e(C_p, C_q)}{\sqrt{\mu|C_p| \cdot \mu|C_q|}} \ .$$

The arithmetic mean of class-level networks is determined by the average class sizes $\mu|C_p|$, $(1 \le p \le k)$ and the average tie weights $\mu\omega(C_p, C_q)$, $(1 \le p \le q \le k)$. It can be regarded as a single instance of a clustered graph (albeit having fractional instead of integer class-sizes) and, thus, can be drawn exactly as in Sect. 3.

4.1.2 Standard Deviation of Clustered Networks

If X is a random variable for which we have N observations $x_1, \ldots, x_N$ and $\mu(x)$ is the mean of these observations, then the *variance* is defined to be $\sigma^2(x) = \sum_{i=1}^{N} (x_i - \mu(x))^2/N$ (i. e., the variance is the average squared difference between observation and mean), and the *standard deviation* is $\sigma(x) = \sqrt{\sigma^2(x)}$.

This formula can be directly used to define the standard deviation of the class sizes

$$\sigma|C_p| = \sqrt{\sum_{i=1}^{N} (|C_p(i)| - \mu|C_p|)^2/N} \ .$$

Similarly to the mean values, the standard deviation of tie-weights is *not* directly obtained from the values $\omega(C_p(1), C_q(1)), \ldots, \omega(C_p(N), C_q(N))$. Instead, we first compute the standard deviation of the un-normalized edge counts

$$\sigma e(C_p, C_q) = \sqrt{\sum_{i=1}^{N} [e(C_p(i), C_q(i)) - \mu e(C_p, C_q)]^2/N} \ ,$$

and normalize these by the geometric mean of the average class sizes to obtain the standard deviation of the tie weights

$$\sigma\omega(C_p, C_q) = \frac{\sigma e(C_p, C_q)}{\sqrt{\mu|C_p| \cdot \mu|C_q|}} \ .$$

The simultaneous visualization of mean and standard deviation is described in Sect. 4.2.

4.1.3 Median and Quartiles

Arithmetic mean and standard deviation have the advantage that they take into account every element of the sample. On the other hand, they have the disadvantage to be heavily influenced by strong outliers. Considering that in quantitative social network analysis strong outliers are very common, we wish to have more stable measures for average and variability. Such measures are, for instance, the median and (upper and lower) quartiles.

If X is a random variable for which we have N observations $x_1, \ldots, x_N$, sorted in non-decreasing order, then the median of these observations is the value that cuts this sequence into two equally-sized halves. More precisely, if N is odd then the median is $x_{\lceil N/2 \rceil}$, if N is even then the median is the arithmetic mean of $x_{N/2}$ and $x_{N/2+1}$. Similarly, the lower quartile is $x_{\lceil N/4 \rceil}$ (except if N is divisible by four, in which case it is the arithmetic mean of $x_{N/4}$ and $x_{N/4+1}$) and the upper quartile is $x_{\lfloor 3N/4 \rfloor}$ (except if N is divisible by four, in which case it is the arithmetic mean of $x_{3N/4}$ and $x_{3N/4+1}$).

If we apply these definitions componentwise to the class-sizes and un-normalized edge counts (and normalize afterwards), we obtain the median and quartiles of clustered networks, just as the mean and deviation (details are omitted).

Obviously, besides mean/deviation and median/quartiles, any other measure for statistical average and variability could be taken.

4.2 Visualization of Average and Variability

To get an idea how average and variability of clustered networks can be visualized, we first have a look at how the popular *box plots* or *box-and-whisker diagrams* [16] achieve this for samples of one-dimensional variables in Fig. 7.

To apply the idea of box plots to visualizations of clustered networks, some adaptions have to be made. First, the networks are multidimensional data sets. More precisely, if we have four classes, then drawings such as Fig. 4 visualize 14 values (four class sizes, four intra-class ties, and six inter-class ties). Clearly, we would like to have all values for mean and deviation in one image. Second, while the value of the variable Y in Fig. 7 is visually encoded in the position on the vertical y-axis, the indicator values for clustered networks are represented differently: class size by the area of the

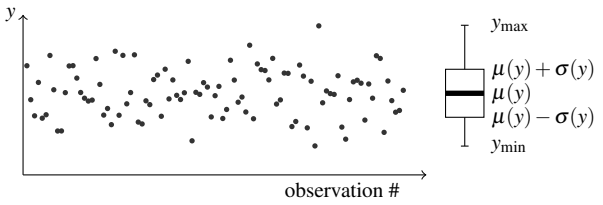


Figure 7: Traditional box plot (*right*) of a sample of a one-dimensional random variable Y. Alternatively, box plots could visualize any other measure for average and variability, e. g.,the median (instead of mean) and upper/lower quartiles (instead of $\mu(y)\pm\sigma(y)$).

circles, intra-class weights by grey-values, inter-class weights by edge width. Taking into account these differences, the idea of box plots can straightforwardly be adapted to images of clustered networks (see Fig. 8 for illustration): The deviation of the size of class C_2 is visualized by drawing a small segment in the lower part of the circle with the radii determined by $\mu|C_2|\pm\sigma|C_2|$. Similarly, the deviation of the inter-class tie between C_1 and C_2 is visualized by increasing/reducing the width on a small part of the corresponding edge by the value determined by $\sigma\omega(C_1,C_2)$. The deviation of the intra-class tie of C_1 is visualized by increasing/reducing the darkness of two wedges in the upper part of the circle by the value determined by $\sigma\omega(C_1,C_1)$. To facilitate the comparison of values, the width of these wedges is proportional to $\sigma\omega(C_1,C_1)$

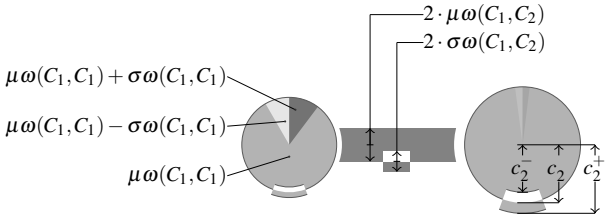


Figure 8: Visual representation of mean and deviation of intra-class weights, inter-class weights, and class sizes for two classes C_1 (*left*) and C_2 (*right*). (Let $c_2=\sqrt{\mu|C_2|}$ denote the square root of the mean size of C_2 and $c_2^{\pm}=\sqrt{\mu|C_2|\pm\sigma|C_2|}$ respectively.). Instead of mean and standard deviation, any other measure for average and variability (e. g., median and upper/lower quartiles) could be taken.

At least in our empirical data set, the observed maximal and minimal values are often close to (or identical with) the feasible maximal and minimal values which are determined by the research design. (For instance, the maximal class size is typically close to 45 and the minimal class size equal to zero.) Since only little information is provided by the feasible extremal values and since the drawings would become vary unbalanced, we do not visualize minimal and maximal values. However, in a scenario where extremal values do not differ so much from the mean/median, these could easily be incorporated into the drawings.

5 Examples

In this section we illustrate how we can get insight into a collection of personal networks by averaging over purposefully chosen subsets. We emphasize once more that the goal of this paper is to propose and illustrate a network visualization technique and not to derive conclusive results about acculturation. Although we detect in the following quite interesting trends, we do not address issues such as significance of statistical observations or representativeness of data sets in this paper. Thus, the results may or may not generalize beyond our given sample.

5.1 Examples for Average and Variability

5.1.1 Dependence on Countries of Origin/Host

First we want to assess how country of origin and host country influence the average acculturation strategy of migrants. To visualize this dependency, we partitioned the whole set of respondents into sub-sets determined by country of origin and target country. We computed the network median and upper/lower quartiles over these sub-samples and visualized them in Fig. 9.

The images in Fig. 9 indeed reveal considerable differences. In the upper row we can see that migrants from Senegal/Gambia have on average many alters originating from and still living in their country of origin and, in addition, actors in this class are tightly connected (dark node color). Most of the alters of the Dominican migrants to Spain belong either to the ORIGIN-class or the FELLOWS-class and—as for the Senegambians—few are originally from the host country. The average networks of Moroccans and Argentineans are more balanced in this respect. In contrast to the Dominicans in Spain, the Dominican migrants to the USA have a much smaller ORIGIN-class. An extremal example in this respect are the Cubans (although we have only seven respondents from this country), whose FELLOW and HOST classes are large and very well connected, whereas the ORIGIN-class is zero. Note that (since we visualize in Fig. 9 the median and quartiles) this does not imply that the ORIGIN-class of all Cuban respondents is zero—just that at most one quarter of the respondents have an alter in this class (in our case it is just one of the seven respondents). A remarkable difference between the upper and lower row is that most migrants to the USA (if we do not consider the seven Cubans) seem to report fewer ties (light-grey colors) in their personal networks than the migrants to Spain (darker colors).

Besides differences in the average we can also observe large differences in the variability. For instance, the Cubans consistently have many alters in their FELLOWS and HOST classes, so that the deviation in the radii is quite small. In contrast, the Dominican migrants to the USA yield a sub-sample with very high variability (so that the median might not well represent many of them). For instance, the lower quartile of the size of their FELLOWS-class is zero and thus very far from the median. Also the tie-weights show different variability: while, e. g., the FELLOWS class of the Haitians shows a small variability in its intra class tie weight (the colors of the two segments are close to the median color), the tie within the FELLOWS of the Dominican migrants to the USA has a high variability (a dark and a bright segment within this node).

5.1.2 Dependence on Time of Residence

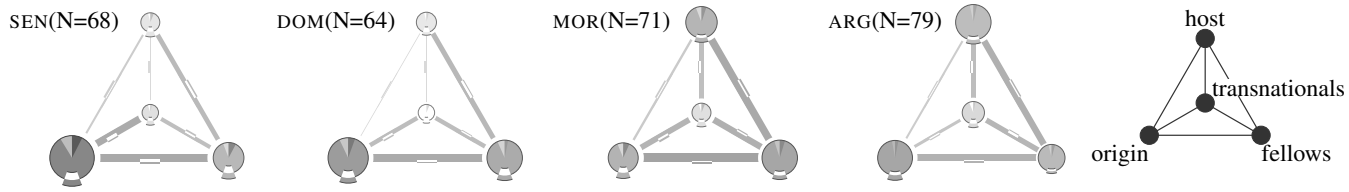
We hypothesize that the level of acculturation also depends on the time of residence. To visualize this, we partitioned the whole set of respondents into sub-sets entering in the same year into their host country and visualize the averages (median and quartiles) over these sub-samples in Fig. 10.

The images in Fig. 10 show an almost monotonic transformation from the country of origin towards the host country. The average migrant seems to start with a social network where most alters are from the ORIGIN-class and form a densely connected community (Y=1). Then the FELLOWS-class and later the HOST-class successively get larger (although not exactly monotonically). In conclusion, Fig. 10 suggests a tendency towards integration over time, although the two ties from HOST to FELLOWS and from FELLOWS to ORIGIN remain important (relatively high weight).

5.2 Refinement of Classes

Until now, all network images in this paper showed the four actor classes defined in Fig. 3. In this section we present some examples of network images with more than just four classes. Exemplary we want to visualize relations between the skin color of respondents and their network composition and structure.

migration to Spain:



migration to the USA:

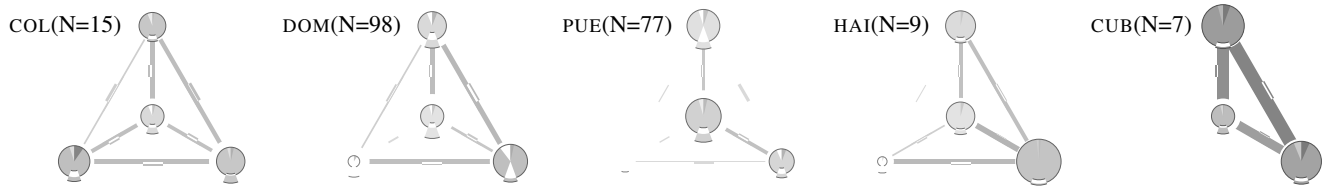


Figure 9: Average networks of migrants with the same country of origin and host country. The network diagram in the upper right position is to remind the positions of the different classes (compare Fig. 3). Upper row are migrants to Spain, lower row are migrants to the USA. The country of origin is encoded as follows: SEN for Senegal/Gambia, DOM for the Dominican Republic, MOR for Morocco, ARG for Argentina, COL for Colombia, PUE for Puerto Rico, HAI for Haiti, and CUB for Cuba,. The integer N denotes the number of individuals in the respective sub-sample.

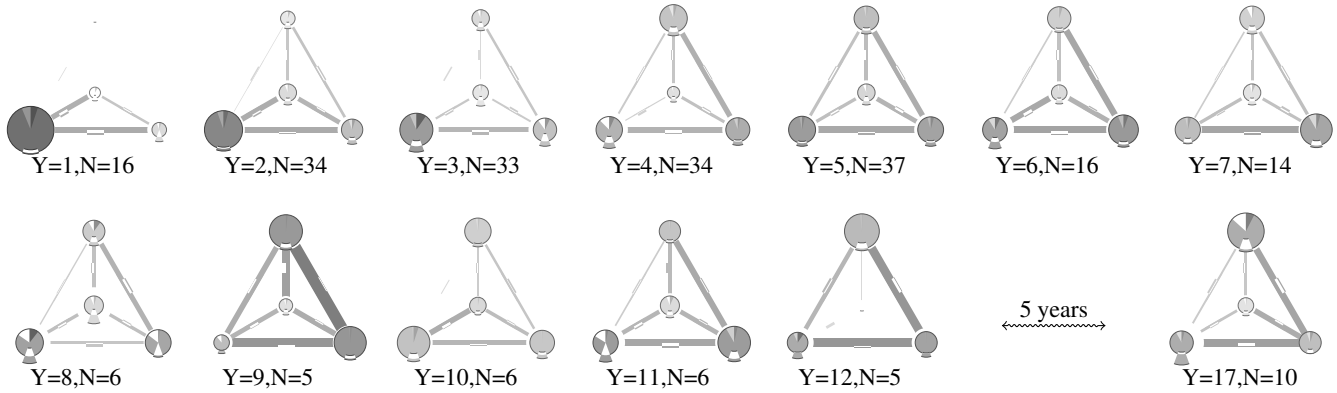


Figure 10: Average networks of migrants with the same year of entry in the host country. The integer Y denotes the years of residence in the host country, the integer N denotes the number of individuals in the respective sub-sample (we only shown networks with $N \geq 5$).

To assess this dependence, we partitioned for each network the four classes (ORIGIN, FELLOWS, HOST, and TRANSNATIONALS) into four subclasses dependent on whether the actors' skin color has been denoted by the respondent as BLACK, BROWN, WHITE, or OTHER. Similarly, we partitioned the set of respondents (i. e., the set of personal networks) into four sub-samples dependent on whether they declared their own skin color as BLACK, BROWN, WHITE, or OTHER. In a first step we computed for each of these sub-samples the arithmetic mean network (as described in Sect. 4) and, in addition, the mean sizes of the "skin color"-subclasses (by exactly the same formulas as we compute the average sizes of classes). The composition of each class is visually encoded in a pie chart (see the top row in Fig. 11).

As for country of origin and time of residence, we can detect considerable dependence on the skin color. On average, BLACK, BROWN, and WHITE respondents know more alters that have their own skin color respectively. (On the other hand, those that declared their own skin color as OTHER seem to have the most multi-racial network.) However, the HOST class does not follow this rule: the largest subclass of the HOST class consists of WHITES, independent on the skin color of the respondent (although the ratio varies). Presumably, this is a characteristic of the two host countries (Spain and the USA) whose populations consist mostly of WHITE people. Furthermore, note that the BLACK respondents have the smallest HOST class, followed by the BROWN, WHITE, and OTHER migrants. This observation leads to the hypothesis that having a different skin color than the largest part of the host society hinders integration. (Although it is unclear why exactly those who declared their skin color as OTHER have the largest HOST class.)

In the bottom row in Fig. 11 we have drawn the networks consisting of the 16 subclasses (ORIGIN, FELLOWS, HOST, and TRANSNATIONALS intersected with the four skin color classes) in a standard circular layout. These images confirm what we observed in the images in the upper row but, in addition, reveal the adjacency structure between subclasses. Although, this structure gets more complicated, some structural properties can be recognized: First, alters having the same skin color seem to be on average better connected, so that dense squares (including diagonals) that connect equally colored nodes become visible. Second, the four sub-classes belonging to the same super-class are often better connected.

The circular layout (as in the bottom row of Fig. 11) could be applied to any number of classes. However, it is evident that choosing too many classes produces images that are too complex. In addition, having too many classes increases the risk of overfitting to the

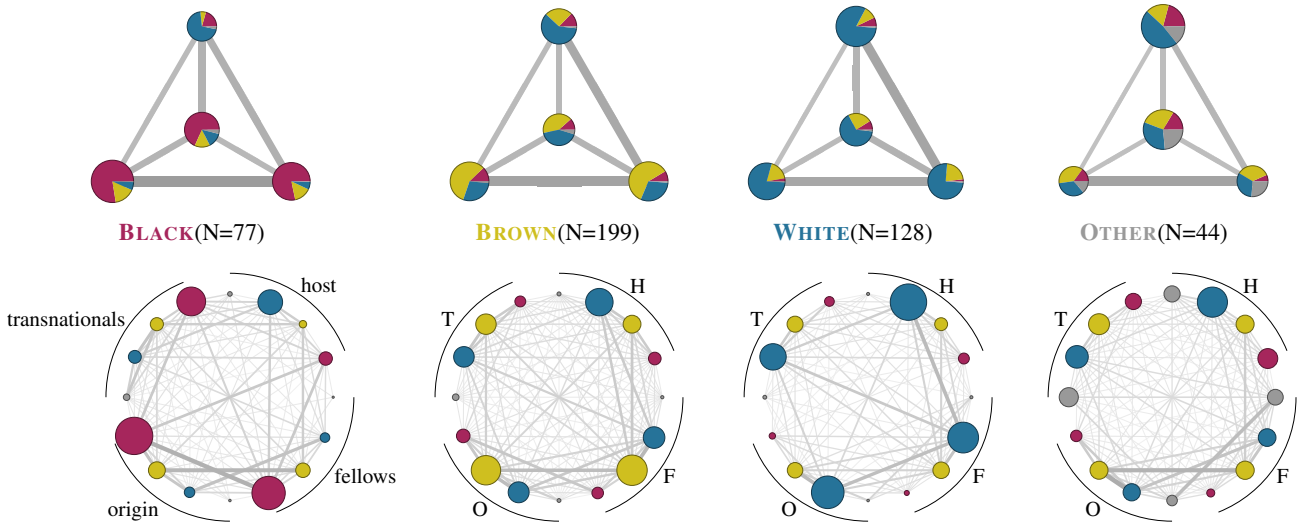


Figure 11: Average networks of migrants (respondents) with the same skin color. The labels in the middle row denote the skin color of the respondent (the color of these labels is used to encode the skin color in the network images), the integer N denotes the number of respondents in the respective sub-sample. In the top row the four classes have been subdivided (as in a pie chart) each into four segments proportional to the relative sizes of sub-classes. The networks in the bottom row show the average adjacency structure of the 16 sub-classes. For simplicity we visualize only the mean and not the deviation in these images.

particular sample, i. e., to obtain observations that do not generalize well to a larger population. A small number of well-chosen classes is likely to yield more useful information to the analyst.

6 Conclusion

We propose techniques for visualizing collections of clustered graphs that may have disjoint vertex sets and where only a one-to-one correspondence between the class labels is given. A distinguishing property of our method is that we abstract from individual vertices and show only the size of vertex classes and how vertices from specific classes are connected on average. In addition, the classes are always displayed at the same position, independent of the specific graph structure. These decisions have been crucial to get small and concise images of graphs, to facilitate easy comparison between disjoint graphs, and to allow for averaging over collections of graphs. Furthermore, we claim that our clustered graph visualizations support the analyst in building models for the generation of ties, i. e., models that predict the probability of ties dependent on the characteristics of the connected actors. We illustrated the usefulness of our method by analyzing a collection of personal networks of migrants. Our images yield visual measures for assessing the mode of acculturation of individuals and show interesting trends in sub-samples, thereby revealing dependency of cultural integration on country of origin, time of residence, and skin color.

A promising direction for future work is to develop techniques to visualize collections of graphs that are hierarchically clustered—and thereby to deal with larger numbers of classes. The subdivision of classes as in a pie-chart (shown in the top row of Fig. 11) is a simple approach, but it does not show the adjacency structure between subclasses. It is likely that other methods for hierarchical clusterings (e. g., [1, 2, 7]) could be adapted to our application.

Acknowledgements

Research supported by DFG grant Br 2158/2-3 and NSF award no. BCS–0417429.

References

[1] M. Balzer and O. Deussen. Level-of-detail visualization of clustered graph layouts. In *Proc. Asia-Pacific Symp. Visualisation*, 2007.

[2] M. Baur and U. Brandes. Multi-circular layout of micro/macro graphs. Proc. 15th Intl. Symp. Graph Drawing, to appear.

[3] J. W. Berry. Immigration, acculturation, and adaptation. *Applied Psychology*, 46(1):5–68, 1997.

[4] U. Brandes and T. Erlebach, editors. *Network Analysis*. Springer Verlag, 2005.

[5] I. Cuéllar, B. Arnold, and R. Maldonado. Acculturation rating scale for Mexican Americans-II: A revision of the original ARSMA scale. *Hispanic Jounal of Behavioral Sciences*, 17(3):275–304, 1995.

[6] G. Di Battista, P. Eades, R. Tamassia, and I. G. Tollis. *Graph Drawing: Algorithms for the Visualization of Graphs*. Prentice Hall, 1999.

[7] P. Eades and Q. Feng. Multilevel visualization of clustered graphs. In *Proc. Intl. Symp. Graph Drawing*, pages 101–112, 1996.

[8] P. Eades and M. L. Huang. Navigating clustered graphs using force-directed methods. *Journal of Graph Algorithms and Applications*, 4(3):157–181, 2000.

[9] Y. Frishman and A. Tal. Dynamic drawing of clustered graphs. In *Proc. IEEE Symp. Information Visualization*, pages 191–198, 2004.

[10] D. Holten. Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data. *IEEE Trans. Visualization and Computer Graphics*, 12(5):741–748, 2006.

[11] M. Kaufmann and D. Wagner, editors. *Drawing Graphs: Methods and Models*. Springer Verlag, 2001.

[12] S. G. Kobourov and C. Pitta. An interactive multi-user system for simultaneous graph drawing. In *Proc. Intl. Symp. Graph Drawing*, pages 492–501, 2004.

[13] A. Noack and C. Lewerentz. A space of layout styles for hierarchical graph models of software systems. In *Proc. ACM Symp. Software Visualization*, pages 155–164, 2005.

[14] R. Redfield, R. Linton, and M. Herskovits. Memorandum on the study of acculturation. *American Anthropologist*, 38:149–152, 1936.

[15] J. F. Rodrigues Jr., A. J. Traina, C. Faloutsos, and C. Traina Jr. SuperGraph visualization. In *Proc. IEEE Symp. Multimedia*, pages 227–234, 2006.

[16] J. W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, 1977.

[17] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.