The Network of Genetic Admixture in Humans

Hend Alrasheed and Feodor F. Dragan

Abstract Recent advances in the field of genetic data analysis reveal promising findings in the field of human history; especially when combined with proper data analysis tools. Within the field of modern genetics, there is evidence that the human populations have genetically interacted as a result of several events. The genetic admixture contains multiple pieces of DNA that have been passed down subsequently through generations making it combine DNA from different source groups. In this paper, we construct and analyze the network of human genetic admixture. We study the topology of this network, we investigate its δ-hyperbolicity (negative curvature), and, using it, identify the core vertices by proposing the δ-hyperbolicity-neighborhood measure that we assign to each vertex.

1 Introduction

Using networks to describe systems that are composed of elements and the interactions or connections between those elements aids analyzing and understanding them. Therefore, networks in multiple disciplines ranging from computer science to systems biology are being modeled as graphs were vertices represent the different elements and edges represent the different interactions among those elements. Within the field of modern genetics, there is evidence that the human populations have interacted throughout history. This interaction, which may occur as a result of migrations, invasions, and slavery, results in transfer of genetics and accordingly creates admixed populations. The genetic admixture contains multiple pieces of DNA that have been passed down subsequently through generations making it combine DNA from different source groups.

The work in [10] uses DNA from many people around the world (95 populations) to identify the mixed source groups and to decide when did those mixing events

H. Alrasheed (✉) · F.F. Dragan
Department of Computer Science, Kent State University, Kent, OH 44242, USA
e-mail: halrashe@kent.edu

F.F. Dragan
e-mail: dragan@cs.kent.edu

H. Cherifi et al. (eds.), *Complex Networks VII*, Studies in Computational Intelligence 644, DOI 10.1007/978-3-319-30569-1_18

had occurred. Their results are presented on an interactive map in [1]. Their work concludes that many populations are results of genetically mixed groups that mixed throughout the last 4000 years. Furthermore, some of those mixed source populations are geographically very spread. Finally, even though genetic mixing among source groups is often local with respect to time and space, neighboring populations do not necessarily share the same ancestry or history.

Even though it is interesting to analyze the details of the direct genetic admix among populations, it is equally interesting to see how this genetic admix looks like in the organization level by the use of graph-theoretical tools. This global approach of analyzing the genetic admix as a system not only as individual components may increase our understanding of the human history in multiple aspects; for example, the transmission of languages and cultures. In this paper, we construct and analyze the network of human genetic admixture. We investigate the topology of this network by studying the degree distribution, the clustering coefficient, and the different measures of centrality. We also investigate the δ-hyperbolicity of this network and, using it, identify the core vertices. For this we propose the δ-hyperbolicity-neighborhood of each vertex. Then we use this measure to identify the core vertices. Based on our analysis, we find the average distance between a pair of populations across the network are relatively small suggesting a small-world network. We also find that the network comprises a number of sub-networks when edges are pruned based on their weights. Those sub-networks are formed by multiple neighboring populations. Also, we identify key vertices according to a number of centrality measures, and we find that those measures correlate very well. Moreover, we find the core vertices identified based on the δ-hyperbolicity-neighborhood measure correlate to some extent with some of the typical centrality measures such as the betweenness centrality.

2 Data and Network Construction

The data was obtained from the *Genetic Atlas of Human Admixture History* interactive website [1], which is a companion of the work presented in [10]. In this work, the authors study 95 populations (a population or a group is a set of individuals with similar genetic makeup). For each individual population p, they show the set of other source populations that are genetically admixed in the DNA of population p. For example, Fig. 1, which is a screen shot from [1], shows that the Polish population has the following admixing groups: Lithuanian (53.1 %), Norwegian (16 %), Russian (12.9 %), Moroccan (3.7 %), Sardinian (3.7 %), Basque (2.6 %), etc. The percent associated with each source group indicates its contribution to that population such that all admixed populations collectively make 100 %. Overall, we found 2685 distinct edges in this network.

Here we construct and study the network of genetic admixture in human populations. In this network vertices represent the different populations and an edge connects two populations if one participates in the genetic makeup of the other. Each edge has an associated direction and a weight. For a source population u and a

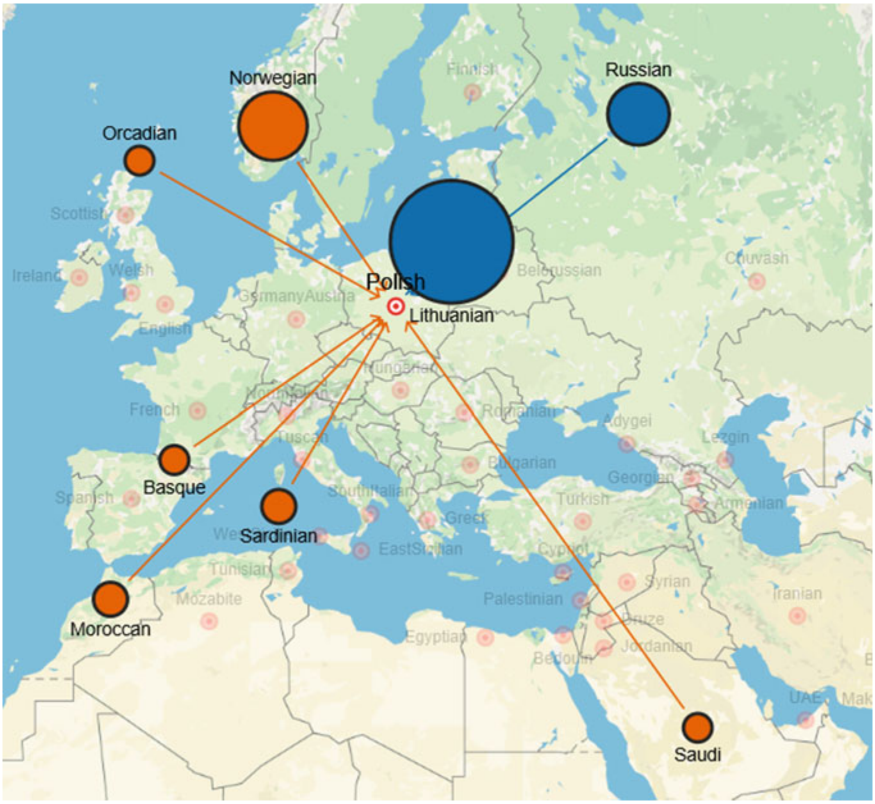

Fig. 1 Genetic mixing in the genetic of the Polish population

population v, the edge e_{uv} is directed towards v. The weight assigned to each edge e_{uv} denoted as w_{uv} is based on the percent of contribution in which it participates in building the DNA of this admixed group. Hence, the larger the weight is, the more significant the contribution. We normalize the weights as follows. The weight of an edge e_{uv} becomes $w_{uv} = (100 - \lambda_{uv})/10$, where λ_{uv} is the percent of the contribution as reported in [1]. This way the smaller the weight is, the larger the contribution, and as a result, the shorter the distance between the two populations. For example, if a source group u represents 50 % of group v's DNA, then the edge leaving u towards v has a weight of 5.

A graph can be expressed by its adjacency matrix a_{uv} where the value a_{uv} is one if vertices u and v are connected and zero otherwise and w_{uv} represents the weight of that edge if one is present. We use this representation throughout this work. For several reasons that will become apparent later on in this text, we will be analyzing the weighted and the unweighted versions of this network. In the weighted network, different edges will have associated weights as described above. In the unweighted network, edges are either present or absent (we ignore w_{uv} in this case). Table 1 gives some overall statistics of the constructed networks.

3 Network Analysis

In this section, we study some fundamental global and local network parameters of the two generated networks: the weighted genetic admixture network and the unweighted genetic admixture network.

Diameter, characteristic path length, and small-world property. According to the distances between vertices in the graph, the *eccentricity* of a vertex u is $ecc(u) = \max_{v \in V}\{d(u, v)\}$. The minimum value of the eccentricity represents the graph's *radius*: $rad(G) = min_{u \in V}\{ecc(u)\}$. The *diameter* of the graph $diam(G)$ refers to the length of the longest shortest path between any two vertices u and v, i.e., $diam(G) = \max_{u,v \in V}\{d(u, v)\}$. Another important distance related measure of graphs is the *characteristic path length* (CPL) which is the average distance between vertex pairs. See Table 1. Many real-world networks exhibit the *small-world property*. A network is said to have this property when it has a small CPL or diameter compared to the size of the network. Let $size(G) = |V| + |E|$ be the size of graph G, a network has the small-world property when $diam(G) \leq \log_2(size(G))$. For our network, $\log_2(size(G)) = 11.44$.

For the unweighted genetic admixture network, the diameter is 4, which is small compared to the network's size. However, since the diameter in graphs is susceptible to outlier vertex pairs [11], we are also interested in the *effective diameter* which represents the maximum distance between a fraction of vertex pairs (in our case 90 %) of the network. The effective diameter for this network is 2, and the CPL is 1.8. Clearly, this network exhibits the small-world property. This indicates that if one population p_1 does not contribute (directly) in the genetic make up of another population p_2, then there is a small chain of population exchanges between the two. For the weighted network, the diameter is ≈36, the effective diameter is ≈19, and the CPL is ≈16. In both networks the diameter is finite which means that all vertices are reachable from one another. In other words, the network of human genetic admixture has one connected component. Also, we find that the network is biconnected.

Weights and network components. One would expect neighboring populations to be genetically admixed; however, it was concluded in [10] that some mixture

Table 1 Basic network parameters

Measure	Directed weighted	Directed unweighted
$diam(G)$	36.38	4
CPL	15.83	1.8
$rad(G)$	19.68	2
$\bar{k}^{+}(G)$	143.8	14.6
$\bar{k}^{-}(G)$	273	28.3
$\bar{k}(G)$	416.8	42.9

$diam(G)$: network's diameter; CPL: characteristic path length; $rad(G)$: graph's radius; $\bar{k}^{+}(G)$, $\bar{k}^{-}(G)$, $\bar{k}(G)$: average in-degree, average out-degree, average total degree respectively events include populations that belong to very distant locations. This is evident considering our genetic admixture network that is represented by one connected component. However, this also motivates investigating the sub-networks that this single component comprises. Specifically, we focus on the number of sub-networks, the size of each sub-network, and how the populations in each sub-network are connected. To obtain the set of sub-networks, we use the edge weights as an indication of their importance. Given a threshold number t, where $0 \leq t \leq 100$, first, we fix the threshold weight t and construct a graph $G^t = (V, E^t)$ by pruning those edges with weights less than t. Second, we identify the set of strongly connected components for the directed network G^t; each strongly connected component represents one sub-network.

We start this process with $t = 100\,\%$ (the highest possible weight). At this point, every single vertex in the graph G^{100} represents a strongly connected component on its own. Then we gradually reduce threshold t (obtaining a different set of sub-networks) until we get to a point in which all vertices are in the same component (when $t = 0$). The number of strongly connected components as well as some other properties about each component are listed in Table 2. An interesting observation about the formation of populations into distinct sub-networks is that it is highly affected by the geographic locations of those populations. For an example, see Table 3 in which we provide a list of all sub-networks with size ≥2 along with the geographic location to which the listed populations belong. The geographic regions are as presented in [10]. Note that all populations in a sub-network either belong to the same geographic region or to a region that is close geographically. This indicates that, for some populations, the genetic admixing with neighboring populations is more significant. Another interesting observation is that the small-world property is evident in the sub-networks. For example, G^3, G^2, and G^1 in Table 2.

Table 2 Sub-networks in each G^t that result from pruning edges with weights $<t$

t (%)	$\lvert G^t \rvert$	$\lvert E^t \rvert$	Min # of vertices in a sub-network	Max # of vertices in a sub-network	Diam of largest sub-network
90	95	2	1	1	0
70	95	7	1	1	0
50	94	18	1	2	1
30	88	63	1	3	1
10	36	222	1	33	8
5	18	418	1	61	5
3	12	601	1	78	6
1	2	1099	5	90	4
0	1	2685	95	95	4

t: edge threshold; $\lvert G^t \rvert$, $\lvert E^t \rvert$: number of sub-networks and edges in G^t; Diam of largest sub-network is the longest (unweighted) path that exists between any two vertices

Table 3 A list of populations in some of the sub-networks in G^t where $t = 20$ and the geographic region(s) of the listed populations

No.	Populations	Geographic region
1	GermanyAustria, Finnish, Norwegian, English, Ireland Scottish, Spanish, French	N.W.Europe, E.Europe
2	Cambodian, Dai, Han, HanNchina, Tujia, Miao	S.EastAsia
3	Balochi, Brahui, Sindhi, Pathan	C.SouthAsia
4	BantuSouthAfrica, SanKhomani, SanNamibia	Bantu, San
5	Belorussian, Polish, Lithuanian	E.Europe
6	Ethiopian, EthiopianJew	Ethiopian
7	BantuKenya, Yoruba	Bantu, W.Africa
8	Adygei, Georgian	W.Asia
9	Bedouin, Saudi	S.MiddleEast
10	Daur, Oroqen	N.EastAsia
11	Yi, Naxi	S.EastAsia

Clustering coefficient. The *clustering coefficient* for a vertex v, denoted as $cc(v)$, indicates the likeliness that any two neighbors of v are also neighbors. Given an unweighted graph $G = (V, E)$ and a vertex $v \in V$, let $N(v)$ be the neighborhood of v consisting of all vertices adjacent to v. Also, let $e_{N(v)}$ be the set of edges between every pair of vertices in v's neighborhood. Then $cc(v) = \frac{2|e_{N(v)}|}{|N(v)||N(v)-1|}$. $0 \leq cc(v) \leq 1$. The clustering coefficient $CC(G) \in [0, 1]$ of a graph G is the average of $cc(v)$ taken over all vertices $v \in V$. $CC(G) = 0$ when there is no clustering and $CC(G) = 1$ when the clustering is very high which happens when the network includes a number of sub-networks each of which is highly dense and connected with other sub-networks with very few links.

For our network, the clustering coefficient measures the tendency of two populations that both already genetically admixed with a third population to themselves admix (we ignore the directions here). The average clustering coefficient of the network is about 0.57. We are also interested in exploring the following: if a population p_1 contributes to the genetic admix of another population p_2, what is the probability that population p_2 also contributes to p_1's admix? This question can be answered using the graph's *reciprocity* which is another important property of directed networks [13]. The reciprocity of a given graph, denoted as $R(G)$, is the fraction of edges that point to both directions (vertices) and it is calculated as $R(G) = \frac{|e^*_{uv}|}{|E|}$, where $|e^*_{uv}|$ is the number of bidirectional edges and $0 \leq R(G) \leq 1$. The reciporcity of our genetic admixture network is 0.24 which means that if population p_1 contributes to the DNA of population p_2, then there is a probability of 24 % that p_2 also contributes to the DNA of population p_1. This could be explained by the one direction immigration. Close analysis of those pairs of populations, that admix in only one direction, shows that the admix involves non-neighboring populations.

Degree distribution and the degree centrality. The *degree* of a vertex u (denoted as k_u) in an undirected graph G is the number of edges that have u as one of their endpoints, i.e., $k_u = \sum_v a_{uv}$. If G is directed, then a vertex u has an *in-degree* denoted as k_u^+ that represents the number of edges in E that have u as a source vertex, and an *out-degree* k_u^- that represents the number of edges that have u as a target vertex. The in-degrees and the out-degrees of the vertices in our directed unweighted genetic admixture network fluctuates between 1 and 63, with Papuan and Druze having the highest in-degree and out-degree respectively. In case of weighted networks, the weights of the edges are important to give a more precise characterization of its complexity. Therefore, rather than considering the number of incident edges, we consider their weights. Hence, the degree k_u is defined as $\sum_v a_{uv} w_{uv}$. The in-degree and the out-degree are defined accordingly. The population with the highest weighted in-degree is Papuan and the population with the highest out-degree is Druze. The *average degree* of the graph G, $\bar{k}(G)$, is defined as $\bar{k}(G) = \frac{1}{|V|} \sum_{u \in V} k_u$. See Table 1.

The degree centrality considers the central vertices as the set of vertices with the highest number of connections. The degree centrality is a local measure since it only relies on the number of neighbors [6]. Therefore, we compute the degree distribution $p(k)$ and the cumulative degree distribution $P(k) = \sum_{\ell \geq k} p(\ell)$ which indicates the fraction of vertices with degree k or larger. The cumulative degree distribution often provides some global characteristic of the network. In Fig. 2, we plot the cumulative in-degree and out-degree distributions for our directed unweighted genetic admixture network in a semilogarithmic scale. One could think that vertices with very high in-degrees act like populations that belong to popular geographical locations that may had attracted immigrants or had represented commercial attractions. However, it also may be the case that the high in-degree is just a result of being geographically close to multiple other populations and the genetic admix is just a consequence of the location.

Distances and centrality. The distance $d(u, v)$ between two vertices u and v in a graph G is the number of edges in a shortest (u, v)-path that connects them. When G is a weighted graph, the distance $d(u, v)$ is the sum of the weights of all edges in a shortest (u, v)-path from u to v (direction is important). The centrality measures presented in this section are all based on the set of shortest paths in a graph. A centrality measure rank the vertices according to their importance. Then it identifies

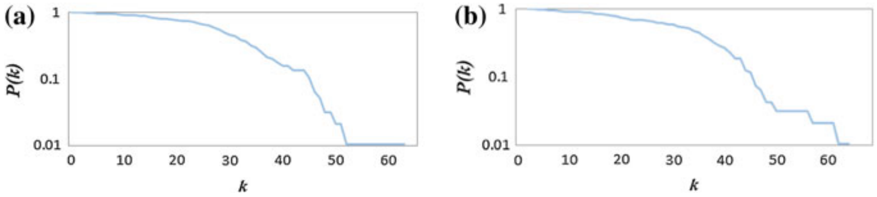

Fig. 2 The cumulative degree distribution $P(k)$ with the in-degree k (**a**) and the out-degree k (**b**). The *horizontal axis* for each chart is the in-degree or the out-degree and the *vertical axis* is the cumulative probability distribution of that degree

the set of vertices that are most significant and accordingly more central. There are multiple centrality measures each of which identifies the key vertices based on a distinct purpose. In this section, we limit our discussion to those measures that are directly based on the notion of distances.

The *betweenness centrality* measure expresses how much effect each vertex has in the communication in the network assuming that all traffic follows shortest paths. Informally, the betweenness centrality of a vertex v refers to the total number of shortest paths between every vertex pair that pass through v. Let $\alpha_{wz}(v)$ be the fraction of shortest paths between w and z that pass through v, i.e., $\alpha_{wz}(v) = \sigma_{wz}(v)/\sigma_{wz}$, where $\sigma_{wz}(v)$ is the number of all shortest paths between w and z that pass through v and σ_{wz} is the number of all shortest paths between w and z. The betweenness centrality $c_B(v)$ of v is $c_B(v) = \sum_{w \in V} \sum_{z \in V} \alpha_{wz}(v)$ [6]. Higher values of this measure indicates higher importance of the vertex. The *closeness centrality* considers the central vertices as the subset of vertices with the minimum total distance to all other vertices. The closeness centrality $c_C(v)$ of a vertex v is defined as $c_C(v) = 1/\sum_{u \in V} d(v, u)$ [6]. The *eccentricity centrality* suggests that the center of the graph includes the vertex (or vertices) that has the shortest distance to all other vertices. For a given vertex v, the eccentricity centrality is $c_E(v) = 1/\max\{d(v, u) : u \in V\}$ [9]. The vertices with the highest eccentricity centrality in fact form the *center* of the network $C(G)$. In other words, $C(G) = \{u \in V : ecc(u) = rad(G)\}$. Tables 4 and 5 list the highest ten populations for the degree, betweenness, eccentricity, and closeness centrality measures for the unweighted and the weighted networks respectively. Note that in the fifth column of Table 4, all the five listed populations have equal eccentricity centrality. For the unweighted network, the Spearman rank correlation coefficient, which tests the association between two sets of ranks, between the betweenness and the closeness centralities is 0.677 with 70 % common populations in the list of top 10 populations. For the weighted network, the Spearman rank correlation between the two measures is about 0.41.

Table 4 Top ten populations with respect to degree, betweenness, eccentricity, and closeness centrality measures for the directed unweighted genetic admixture network

In-degree	Out-degree	Tot-degree	Betweenness	Eccentricity	Closeness
Papuan	Druze	Adygie	Burusho	Papuan	Druze
Maya	Palestinian	Armenian	Papuan	Melanesian	Palestinian
Melanesian	Burusho	Balochi	Druze	Columbian	Burusho
Burusho	Maya	BantuKenya	Melanesian	Lahu	Maya
Uzbekistani	Hazara	BantuSouthAfrica	IndianJew	Hazara	Hazara
IndianJew	Melanesian	Basque	Maya	–	Melanesian
Cambodian	Sardinian	Bedouin	Hazara	–	Papuan
Adygie	Papuan	Belorussian	Palestinian	–	Sardinian
Turkish	Kalash	BiakaPygmy	Adygie	–	Kalash
Pathan	Indian	Brahui	MbutiPygmy	–	Brahui

Table 5 Top ten populations with respect to degree, betweenness, eccentricity, and closeness centrality measures for the directed weighted genetic admixture network

In-degree	Out-degree	Tot-degree	Betweenness	Eccentricity	Closeness
Papuan	Druze	Papuan	Spanish	Burusho	Druze
Melanesian	Palestinian	Burusho	Maya	Armenian	Han
Indian	Burusho	Maya	Han	Pathan	Palestinian
Burusho	Maya	Melanesian	SanKhomani	Maya	Maya
Maya	Hazara	Indian	Moroccan	Hazara	Burusho
Pathan	Indian	Palestinian	Iranian	Melanesian	Balochi
Myanmar	Melanesian	Druze	Cypriot	Papuan	Jordanian
Cambodian	Papuan	Hazara	Pathan	She	Moroccan
IndianJew	Bedouin	IndianJew	Adygei	Han	Brahui
Adygei	Mozabite	MbutiPygmy	EastSicilian	Yakut	Melanesian

4 δ-Hyperbolicity and Network's Core

δ-Hyperbolicity is a measure that captures the notion of negative curvature in abstract metric spaces including graphs. A simple graph $G = (V, E)$ naturally defines a metric space (V, d) on its vertex set V where the distance $d(u, v)$ is defined as the length a shortest (v, u)-path between v and u. In graphs, δ-hyperbolicity measures how close the graph's structure is to a tree structure metrically [8]. Given a graph $G = (V, E)$, x, y, u, and $v \in V$ are four distinct vertices, and the three sums: $d(x, y) + d(u, v)$, $d(x, u) + d(y, v)$, and $d(x, v) + d(y, u)$ sorted in a non-increasing order, the hyperbolicity of the quadruple x, y, u, v denoted as $\delta(x, y, u, v)$ is defined as: $\delta(x, y, u, v) = (d(x, y) + d(u, v)) - (d(x, u) + d(y, v))/2$. The δ-hyperbolicity of the graph is $\delta(G) = \max_{x,y,u,v \in G} \delta(x, y, u, v)$. Generally, the smaller the value of $\delta(G)$ the closer the graph is to a tree metrically and, as a result, the hyperbolicity property is more evident. Even though the δ-hyperbolicity by definition considers the maximum difference between any two largest distance sums for any quadruple, recent research also analyzes the distribution of δ-hyperbolicity of the quadruples [2, 3, 7]. This makes the value of the average δ-hyperbolicity (taken over all quadruples) equally important. The small δ-hyperbolicity property has been found in many real-world networks [2, 3, 7, 12]. However, in many of those networks, this low value is a direct result of the small-world property especially that the inequality $\delta(G) \leq \frac{diam(G)}{2}$ is sharp. For our unweighted network, $\delta(G) = 2$ and the average δ-hyperbolicity of the graph $\delta'(G)$ is 0.24. For our weighted network, $\delta(G) = 15$ and $\delta'(G) = 1.96$.

δ-Hyperbolicity, centrality, and network's core. It was suggested in [4] that the concentration of load on a subset of vertices of the network, for communication assuming shortest path routing, is due to its negative curvature or δ-hyperbolicity. This concentration can be seen as a bend in those shortest paths towards a core of the network defined by its most central vertices. However, the identification of core vertices differ according to the centrality measure used to decide the central vertices. In [4], the core is defined as the subset of vertices with highest betweenness centrality. In [3] the core is defined based on the eccentricity centrality and the betweenness centrality.

Proposition 1 ([3]) *Let G be a δ-hyperbolic graph and x, y be arbitrary vertices of G. If $d(x, y) > 4\delta + 1$, then on any shortest (x, y)-path there is a vertex w with $ecc(w) < \max\{ecc(x), ecc(y)\}$.*

According to the proposition, shortest paths bend towards vertices with smaller eccentricity making the graph's core mostly represented with vertices with the smallest eccentricity ($rad(G)$ or $rad(G) + 1$ in most cases). Then, the betweenness centrality is used to prioritize those vertices according to their participation. In [2], it has been observed that if one constructs a small r-neighborhood where ($r = \delta(G)$) around a vertex v on a shortest path between two vertices x and y, then all shortest paths between x and y include a vertex in this r-neighborhood.[1] Our goal is to identify the core vertices using the δ-hyperbolicity of the network without any presumptions about the centrality of the vertices in the network. Then we analyze the core vertices in terms of their centrality.

For each integer $r \geq 0$, let $N_r(u)$ denotes the *neighborhood* of distance at most r centered at u, i.e., $N_r(u) = \{v \in V : d(u, v) \leq r\}$. We define the δ-hyperbolicity-neighborhood of a vertex u, denoted as $N_\Delta(u)$, as the smallest integer Δ, where $0 \leq \Delta \leq \delta(G)$, such that the majority of vertex pairs (more than 90 %) are covered by that neighborhood. We say a vertex pair (w, z) is covered by the δ-hyperbolicity-neighborhood of a vertex v, if there is at least one vertex $u \in N_\Delta(v)$ such that $d(w, z) = d(w, v) + d(v, z)$. Figure 3 shows that the δ-hyperbolicity-neighborhoods of the majority of vertices when $\Delta = 0$ cover a small percent of vertex pairs (between 3 % and 15 %). An exception is those vertices with higher betweenness; for example, Papuan that covers about 33 % of other vertex pairs. However, when $\Delta = 1$, the δ-hyperbolicity-neighborhood around each vertex covers the majority of vertex pairs. Again some exceptions include French that covers only 27 %. For the details, take a look at Table 6.

[1] Note that the value of r could be higher but never exceeds $6\delta(G) + 2$ [2]. However, for real-networks it was observed in [2] that $r \approx \delta(G)$.

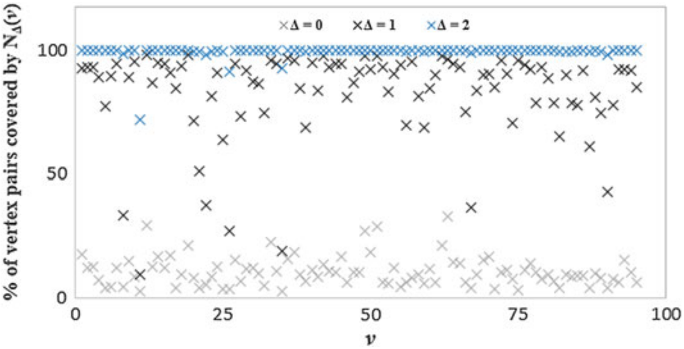

Fig. 3 Percent of vertex pairs covered by the δ-hyperbolicity-neighborhood $N_{\Delta}(v)$ of each vertex v in the directed unweighted genetic admixture network. $1 \le v \le 95$ and $0 \le \Delta \le \delta(G) = 2$

Table 6 Percent of vertex pairs covered by each δ-hyperbolicity-neighborhood in the directed unweighted genetic admixture network

v	Population	$N_0(v)$ (%)	$N_1(v)$ (%)	$N_2(v)$ (%)
a Populations with high coverage				
12	Burusho	29.48	98.53	100
33	Hazara	22.5	96.2	100
63	Papuan	33.08	96.77	100
b Populations with low coverage				
10	Bulgarian	2.8	9.18	72.06
25	French	3.68	27.22	91.58
35	Hungarian	2.79	18.8	93.1

Now we can rank our vertices according to their δ-hyperbolicity-neighborhoods. Each vertex v has two values: (1) Δ, that represents the smallest integer $\Delta \le \delta(G)$ such that the δ-hyperbolicity-neighborhood $N_{\Delta}(v)$ covers more than 90 % of vertex pairs, and (2) the percent of vertex pairs covered by this δ-hyperbolicity-neighborhood. We lexicographically sort all vertices according to those two values. The results are listed in Table 7. The higher the ranking of a vertex, the more it becomes part of the core set.

Discussions. From Fig. 3 and the results listed in Table 7 it is clear that the ranking of vertices obtained according by the coverage of their δ-hyperbolicity-neighborhoods corresponds to some extent with the ranking obtained from the centrality measures; especially the out-degree centrality, the betweenness centrality, the eccentricity centrality, and the closeness centrality. One can see that the majority of the top ten populations in each centrality measure are also present in the top ten list of the vertices with respect to their δ-hyperbolicity-neighborhoods. In contrast, some populations who are not at the top of the ranking with respect to some centrality measures

Table 7 Top ten populations with respect to the δ-hyperbolicity-neighborhoods of vertices in the directed unweighted genetic admixture network

Rank	Population	In-degree rank	Out-degree rank	Betweenness rank	Eccentricity rank	Closeness rank
1	Burusho	4	3	1	2	3
2	Druze	19	1	3	2	1
3	Palestinian	16	2	8	2	2
4	Melanesian	3	5	4	1	6
5	Kalash	17	7	19	2	9
6	Maya	2	4	6	2	4
7	Indian	10	7	15	2	11
8	Papuan	1	7	2	1	7
9	IndianJew	5	9	5	2	22
10	Sardinian	29	6	35	2	8

Here we compare this rank of each vertex with its rank according to the five centrality measures discussed earlier: the in-degree, out-degree, betweenness, eccentricity, and closeness

actually appear as core vertices according to their δ-hyperbolicity-neighborhood. For example, the three populations: Kalash, Indian, and Sardinian all are considered as core vertices according to their δ-hyperbolicity-neighborhood; however, they have lower values for the eccentricity centrality and/or the betweenness centrality measures. This can be justified by the existence of multiple core vertices distributed over multiple cores of the network defined using different centrality measures (or even by the existence of a number of nested cores). Some core vertices are more important with respect to their location and according to the percent of other vertex pairs they cover. This makes those vertices have higher values for the eccentricity centrality and/or the betweenness centrality measures. Still other core vertices, which may have lower eccentricity or betweenness centralities, are important (i.e., essential for communication) for a smaller percent of vertices. This observation motivates investigating the existence of multiple communities that revolve around those different core vertices. Generally, communities in a network are represented by a number of highly dense (with respect to the number of connections) set of vertices; and different communities are linked with fewer connections. Here we use the Louvain method [5] for detecting communities in our unweighted genetic admixture network (we ignore the direction of the edges here). The method identifies three communities (or modules) in our network which admits a modularity of 1.62. The modularity here measures the density of connections inside communities to the density of connections outside communities. See Figs. 4 and 5. Unlike the sub-networks identified earlier in Sect. 3, the modules are represented mostly by non-neighboring populations, and the core vertices are distributed among the different modules.

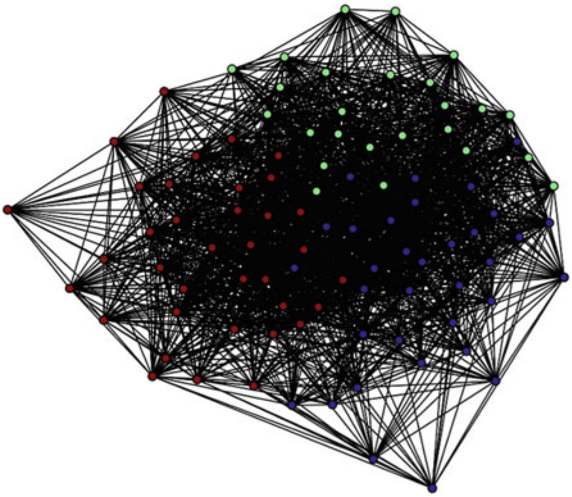

Fig. 4 Populations in three different modules. Module sizes (with respect to the number of vertices) are 36, 34, 25

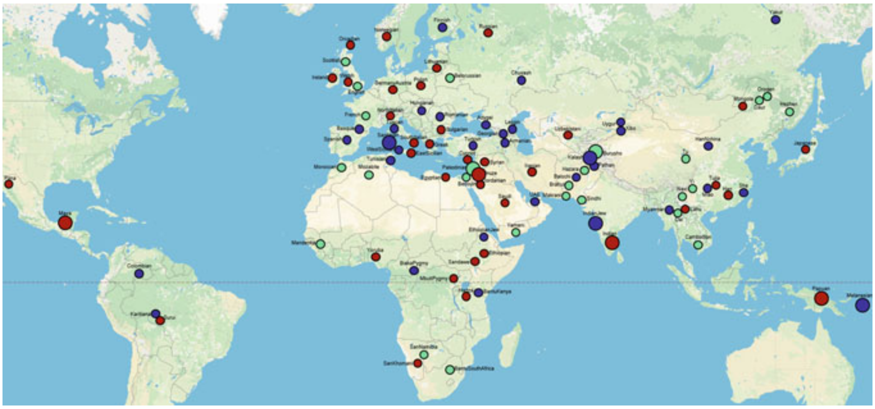

Fig. 5 Each population is assigned to a different module. *Larger circles* indicates the core vertices in each module based on the δ-hyperbolicity-neighborhood of vertices

5 Conclusion

We have studied the genetic admixture network of humans in which vertices represent populations and a link exists between a pair of populations if one participates in the genetic admix of the other. We have considered both the weighted and the unweighted versions of the network. The networks studied were based on data published in [10]. Based on our analysis, we find the average distance between a pair of populations across the network are relatively small suggesting a small-world network. We also find that the network comprises a number of sub-networks when edges are pruned based on their weights. Those sub-networks are formed by multiple neighboring populations. Also, we identify key vertices according to a number of centrality measures, and we find that those measures correlate very well. Finally, we propose a method of identifying core vertices based on the δ-hyperbolicity of the network. This network is dynamic; i.e., the connections among populations are based on a specific time frame. It is interesting to capture different admixing statistics based on various time frames and compare how the dynamicity of this network changes with time.

References

1. http://www.admixturemap.paintmychromosomes.com
2. Albert, R., DasGupta, B., Mobasheri, N.: Topological implications of negative curvature for biological and social networks. Phys. Rev. E **89**(3), 032811 (2014)
3. Alrasheed, H., Dragan, F.F.: Core-periphery models for graphs based on their δ-hyperbolicity: an example using biological networks. In: Complex Networks VI, pp. 65–77. Springer (2015)
4. Baryshnikov, Y.: On the curvature of the internet. In: Workshop on Stochastic Geometry and Teletraffic. Eindhoven, The Netherlands (2002)
5. Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech: Theory Exp. **2008**(10), P10008 (2008)
6. Brandes, U., Erlebach, T.: Network analysis: methodological foundations, vol. 3418. Springer Science & Business Media (2005)
7. De Montgolfier, F., Soto, M., Viennot, L.: Treewidth and hyperbolicity of the internet. In: 2011 10th IEEE International Symposium on Network Computing and Applications (NCA), pp. 25–32. IEEE (2011)
8. Gromov, M.: Hyperbolic Groups. Springer (1987)
9. Hage, P., Harary, F.: Eccentricity and centrality in networks. Soc. Netw. **17**(1), 57–63 (1995)
10. Hellenthal, G., Busby, G.B., Band, G., Wilson, J.F., Capelli, C., Falush, D., Myers, S.: A genetic atlas of human admixture history. Science **343**(6172), 747–751 (2014)
11. Leskovec, J., Chakrabarti, D., Kleinberg, J., Faloutsos, C.: Realistic, mathematically tractable graph generation and evolution, using kronecker multiplication. In: Knowledge Discovery in Databases: PKDD 2005, pp. 133–145. Springer (2005)
12. Narayan, O., Saniee, I.: Large-scale curvature of networks. Phys. Rev. E **84**(6), 066108 (2011)
13. Newman, M.E., Forrest, S., Balthrop, J.: Email networks and the spread of computer viruses. Phys. Rev. E **66**(3), 035101 (2002)