STATISTICAL INFERENCE ON RANDOM PARTITIONS[1]

Lídia Rejtő,[a,b] and Gábor Tusnády[a]

[a] *Alfréd Rényi Mathematical Institute of the Hungarian Academy of Sciences, Budapest, P.O. Box 127, H-1364, Hungary* E-mail address: rejto@renyi.hu, tusnady@renyi.hu

[b] *University of Delaware, Statistics Program, 214 Townsend Hall, Newark, DE 19717-1303, USA* E-mail address: rejto@udel.edu

Budapest, July 16, 2006

[1]Discussants notice to the talks of S. Pentland: "Mapping Social Networks" and of J.T. Rigby: "The Elucidation of Two Mode Block Model Structure Via Spectral-based Co-clustering", held on **Social Networks and Complexity – Workshop**, Collegium Budapest July 31–August 2, 2006

Once upon a time there was a party with 14 participants labeled by integers from 1 to 14. As it is usual in parties they formed groups which were sensed and recorded by devices offered by our modern technics. The data (SIRP DATA) can be found on the home page of G. Tusnády (http://www.renyi.hu/ tusnady/level.dat). The first part is the following

Time	**Participants**													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0.019238	1	1	1	4	5	6	5	6	5	5	5	6	13	5
0.064438	1	1	1	4	5	6	5	6	5	5	5	6	4	5
0.107385	1	1	1	4	5	6	5	6	5	5	5	6	13	5
0.119281	1	1	1	4	5	6	7	6	5	7	7	6	13	7
0.127421	1	1	1	4	5	5	7	5	5	7	7	5	13	7
0.159595	1	1	1	4	5	6	7	6	5	7	7	5	13	7
0.244247	1	1	1	4	5	6	1	6	5	1	1	5	13	1
0.246863	1	1	1	4	5	4	1	4	5	1	1	5	13	1
0.393910	1	2	2	4	5	4	1	4	5	1	1	5	13	1
0.466802	1	1	1	4	5	4	1	4	5	1	1	5	13	1
0.506604	1	1	1	4	5	4	7	4	5	7	7	5	13	7
0.518243	1	1	1	4	5	6	7	6	5	7	7	5	13	7
0.519593	1	1	1	4	5	4	7	4	5	7	7	5	13	7
0.576503	1	1	1	4	5	4	5	4	5	5	5	5	13	5
0.707155	1	1	1	4	5	4	5	4	5	5	11	5	13	5
0.716638	1	1	1	4	5	6	5	6	5	5	11	5	13	5
0.727348	1	1	1	4	5	6	5	6	5	5	11	5	11	5
0.733247	1	1	1	4	5	4	5	4	5	5	11	5	11	5
0.834109	1	1	1	4	5	6	5	6	5	5	11	5	11	5
0.900159	1	1	1	4	5	5	5	5	5	5	11	5	11	5
0.918424	1	1	1	4	5	6	5	6	5	5	11	5	11	5
0.998953	1	1	1	4	5	6	5	6	5	5	11	5	13	5
1.155627	1	1	1	4	5	1	5	1	5	5	11	5	13	5
1.252516	1	1	1	4	5	6	5	6	5	5	11	5	13	5

Each record of the data represents one grouping (partition) formed in the course of the party. The first number means the time when the actual grouping occurred and the next 14 integers denote the partition. Each set of a partition is labeled by its smallest member what we call leading member. For example

Time	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0.900159	1	1	1	4	5	5	5	5	5	5	11	5	11	5

means that after 0.9001594 seconds from the beginning of the party the following groups were sensed by our detectors:

$1+2+3$,
4,
$5+6+7+8+9+10+12+14$,
$11+13$.
Poor 4, seemingly a lonely person walked alone, the noisy central body $5+6+7+8+9+10+12+14$ was situated around the dinner table, while $1+2+3$ had a very important discussion in a secret corner and $11+13$ were playing table tennis.

Performing any data mining the first step is the construction of appropriate colored pictures. In our figures the persons are represented by colors:
1 = red, 2 = yellow, 3 = green, 4 =turquoise, 5 = blue, 6 = white, 7 = purple, 8 =vintage red, 9 = brown, 10 =dark green, 11 = silver, 12 = black, 13 =dark blue, 14 =dark purple, and the partitions are represented by "flowers". Each person is represented as a petal of the flower having a fixed position. But the color of the person is changing according to the partition. Each petal shares the color of the leading member of the group. The color of the center of the flower represents the color of the largest group. Figure 1 shows the bird's view of the occurred partitions where the position of each flower is the result of a **contractive algorithm**. (In Figure 1 flowers are represented by their center only.)

In multidimensional data analysis it is a general idea to join objects whenever they have something common. The trouble is that without any constrains the idea leads to one single point. We use multidimensional covariance standardization as a constraint: it means to centralize the data by averaging them and dividing by the standard deviation and using covariances to enforce the dimension in the required size. The effect resembles to opening an umbrella: the wires spread out what the canopy pulls together.

In a good party there are appropriate places for people willing to do something jointly. But to ensure for all subgroups some special position is prohibitive for combinatorial explosion. We restrict our algorithm to pairs: all pairs of the partitions have same special meeting point and the groups join come together at the average point of their pairs. The formal description of the algorithm is the following.

Let N be the number of different partitions occurred in the party and let x_k, y_k be the coordinates of the point representing the k-th partition ($k = 1, 2, .., N$). The initial values of the coordinates are random standard normal numbers. The iteration consists of the following steps:

Step 1. **Opening the umbrella** (Schmidt orthogonalization):

$$\tilde{x}_k = \frac{x_k - \overline{x}}{w(x)},$$

where

$$\overline{x} = \frac{1}{N}\sum_{k=1}^{N} x_k$$

and

$$w(x) = \sqrt{\sum_{k=1}^{N}(x_k - \overline{x})^2}$$

$$\tilde{y}_k = \frac{y_k^* - c(xy) * x_k}{w(y)},$$

where

$$y_k^* = y_k - \overline{y}$$

$$\overline{y} = \frac{1}{N}\sum_{k=1}^{N} y_k$$

$$c(xy) = \sum_{k=1}^{N} y_k^* * x_k$$

$$w(y) = \sqrt{\sum_{k=1}^{N}(y_k^* - c(xy) * x_k)^2}.$$

Step 2. Positioning pairs of persons (averaging the partitions where the pair happens to be in the same group)

$$u_{i,j} = \frac{\sum_{k:p(i,k)=p(j,k)} \tilde{x}_k * t_k}{\sum_{k:p(i,k)=p(j,k)} t_k}$$

$$v_{i,j} = \frac{\sum_{k:p(i,k)=p(j,k)} \tilde{y}_k * t_k}{\sum_{k:p(i,k)=p(j,k)} t_k},$$

where $p(i,k)$ denotes the leading person in the k-th partition of the i-th person and t_k is the duration of the k-th partition.

Step 3. Dynamics (relocating the partitions with the gradient of the pairs they unite)

$$x_k^{\text{new}} = \tilde{x}_k - \gamma * xx_k$$

where γ denotes a small positive constant (it controls the speed of the algorithm) and

$$xx_k = \sum_{i,j:p(i,k)=p(j,k)} u_{i,j},$$

$$y_k^{\text{new}} = \tilde{y}_k - \gamma * yy_k$$

where

$$yy_k = \sum_{i,j:p(i,k)=p(j,k)} v_{i,j}.$$

Figure 1 shows the partitions and the red rectangle refers to Figure 3. Figure 2 gives the flowers where the system spent more then 20000000 seconds. Figure 3 shows the flowers situated in the red rectangle of Figure 1, where the system spent more then 50000 seconds.

Our data were generated by the distribution

$$P_A(\pi) = \frac{1}{\Gamma(A)} \exp(Q(\pi, A)),$$

where the potential $Q(\pi, A)$ was defined by

$$Q(\pi, A) = \sum_{1 \le i < j \le c:\, \pi(i) = \pi(j)} a_{i,j}.$$

$$\Gamma(A) = \sum \exp(Q(\pi, A))$$

is the scaling factor where the summation runs over all partitions π. The matrix $A = a_{i,j}$ is given by

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	0.00	-3.08	-0.40	0.23	0.75	-1.23	1.13	-0.02	0.51	-2.62	0.78	2.93	0.05	-2.66
2	-3.08	0.00	-4.63	1.18	1.43	1.19	4.73	-0.59	2.61	-5.01	2.08	1.94	3.91	1.94
3	-0.40	-4.63	0.00	4.31	1.65	2.94	-1.17	0.87	-1.37	-0.72	1.66	-0.92	0.19	0.01
4	0.23	1.18	4.31	0.00	2.15	-1.10	-2.77	1.63	4.13	6.79	6.29	1.51	0.41	2.92
5	0.75	1.43	1.65	2.15	0.00	-0.06	-2.31	-0.35	-1.52	-2.30	2.48	-2.23	2.25	-0.72
6	-1.23	1.19	2.94	-1.10	-0.06	0.00	2.75	-4.47	1.29	-4.41	0.38	-4.07	3.75	2.54
7	1.13	4.73	-1.17	-2.77	-2.31	2.75	0.00	-2.86	-2.50	-6.20	-3.35	0.30	3.98	-0.68
8	-0.02	-0.59	0.87	1.63	-0.35	-4.47	-2.86	0.00	-1.16	3.12	-0.27	3.61	5.84	2.25
9	0.51	2.61	-1.37	4.13	-1.52	1.29	-2.50	-1.16	0.00	3.29	5.46	-1.05	1.64	-7.29
10	-2.62	-5.01	-0.72	6.79	-2.30	-4.41	-6.20	3.12	3.29	0.00	-5.14	-2.20	-1.12	-1.77
11	0.78	2.08	1.66	6.29	2.48	0.38	-3.35	-0.27	5.46	-5.14	0.00	5.45	-0.32	-4.20
12	2.93	1.94	-0.92	1.51	-2.23	-4.07	0.30	3.61	-1.05	-2.20	5.45	0.00	3.07	1.74
13	0.05	3.91	0.19	0.41	2.25	3.75	3.98	5.84	1.64	-1.12	-0.32	3.07	0.00	2.82
14	-2.66	1.94	0.01	2.92	-0.72	2.54	-0.68	2.25	-7.29	-1.77	-4.20	1.74	2.82	0.00

The maximum of the potential $Q(\pi, A)$ is 39.64 and it is achieved for the partition

$$\pi = \{1,\ 1,\ 1,\ 4,\ 5,\ 6,\ 5,\ 6,\ 5,\ 5,\ 5,\ 6,\ 13,\ 5\}.$$

The distribution can be sampled by the Metropolis algorithm ([10]). A detail of the process is given in Figure 4, where each horizontal line represents one partition with its color coding. The top red curve shows the change of the potential. Figure 5 gives a time segments of the process on the flowers meadow. The Metropolis algorithm is based on the graph where the vertices are the partitions. We say that two partitions are connected by an edge whenever one is formed by the other with merging of two of its groups. The price of the edge is the product

of the numbers of persons in the amalgamated groups. The distance of two arbitrary vertices is the cheapest path between them

$$d(\pi_1, \pi_2) = \sum_{i=1}^{c} \sum_{j=1}^{c-1} \sum_{k=j+1}^{c} (\nu(i,j) * \nu(i,k) + \nu(j,i) * \nu(k,i)),$$

where

$$\nu(i,j) = \sum_{k=1}^{c} \{\pi_1(k) = i, \pi_2(k) = j\},$$

and c is the number of persons ($c = 14$). In Figure 5 the flowers are positioned by the `isoMDS` multidimensional scaling program of the package R, using the above defined distances. Figures 1,2,3 and 5 are showing the 720 different partitions with potential larger than 35.00. Figure 6 shows 14992 different partitions with potential larger than 28.00 positioned by the contractive algorithm.

One can prove that

$$E_A Q(\pi, B) \quad = \quad \frac{\Gamma(B)}{\Gamma(A)}$$

and

$$\Gamma(A) \leq \prod_{i=1}^{c-1} \prod_{j=i+1}^{c} (1 + \exp(a_{i,j})),$$

but we do not have an explicit form for $\Gamma(A)$. We generated the matrix A as random Gaussian number with zero expectation and standard deviation 2.5 thus the model has the flavor of spin glass processes: there is an abundance of local maxima of the potential and the process is spending the majority of the time in the potential valleys with short time jumping between them ([12]). This might be the case with real world parties where the different partitions are evaluated by the well-being of the persons inside the actual groups. Our model is the simplest possible one because it is based on pair-relations only. Any generalization for higher order interactions is straightforward. The pair-potentials $a_{i,j}$ can be estimated by the maximum likelihood equation ([1]):

$$\frac{1}{N} \sum_{k:p(i,k)=p(j,k)} 1 \quad = \quad P_A(\pi(i) = \pi(j)) \qquad 1 \leq i < j \leq c,$$

or by simple linear regression. The estimated potentials are given in Figure 7, seemingly maximum likelihood is more effective and in both cases estimators of small values have larger error.

Partitions occur in Szemerédi's famous Regularity Lemma ([15]), which was extended recently to hypergraphs ([8], [9]). Parties and hypergraphs are appeared as early as 1941 in the literature [7] (cited in [5]) where 18 ladies attending on 14 parties are investigated. It is a special case of partitions when only two partitions is considered, whether each person present or absent in a party. For a statistician Szemerédi's regularity lemma and it's generalizations

state that the random graph respectively the hypergraph model is bold enough. It may happen that our random model on partitions has the same property. The original proof of Szemerédi is based on an evaluation of the graph and a partition of its vertices. With the help of a duality argument our conjecture is that for any collection of partitions of a large enough set, there exists a graph on the same set, which behaves randomly with respect to the given partitions. For people investigating networks the Albert-Barabási preference model is the most popular one ([11]). Unfortunately it results in graph with moderate number of edges. Any unification of ideas of the models would be welcomed.

The probabilities $P_A(\pi(i) = \pi(j))$ are given in the next table.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	0.0000	0.7548	0.7170	0.0167	0.0744	0.0850	0.2389	0.1071	0.1100	0.4139	0.3203	0.0229	0.0209	0.2941
2	0.7548	0.0000	0.9440	0.0046	0.0053	0.0497	0.0798	0.0836	0.0341	0.2547	0.1616	0.0327	0.0087	0.1235
3	0.7170	0.9440	0.0000	0.0009	0.0239	0.0285	0.1004	0.0653	0.0658	0.2489	0.1552	0.0537	0.0142	0.1478
4	0.0167	0.0046	0.0009	0.0000	0.0040	0.1409	0.0099	0.0868	0.0006	0.0000	0.0001	0.0759	0.3026	0.0000
5	0.0744	0.0053	0.0239	0.0040	0.0000	0.2797	0.5433	0.2885	0.6778	0.4168	0.3449	0.4957	0.0013	0.5340
6	0.0850	0.0497	0.0285	0.1409	0.2797	0.0000	0.1612	0.6676	0.1404	0.2063	0.1486	0.5505	0.0002	0.0136
7	0.2389	0.0798	0.1004	0.0099	0.5433	0.1612	0.0000	0.2792	0.4541	0.7954	0.7794	0.1287	0.0031	0.6856
8	0.1071	0.0836	0.0653	0.0868	0.2885	0.6676	0.2792	0.0000	0.2581	0.1678	0.2111	0.2466	0.0001	0.1010
9	0.1100	0.0341	0.0658	0.0006	0.6778	0.1404	0.4541	0.2581	0.0000	0.2809	0.2705	0.3658	0.0052	0.6779
10	0.4139	0.2547	0.2489	0.0000	0.4168	0.2063	0.7954	0.1678	0.2809	0.0000	0.8192	0.1219	0.0090	0.5958
11	0.3203	0.1616	0.1552	0.0001	0.3449	0.1486	0.7794	0.2111	0.2705	0.8192	0.0000	0.0206	0.0571	0.5911
12	0.0229	0.0327	0.0537	0.0759	0.4957	0.5505	0.1287	0.2466	0.3658	0.1219	0.0206	0.0000	0.0042	0.1571
13	0.0209	0.0087	0.0142	0.3026	0.0013	0.0002	0.0031	0.0001	0.0052	0.0090	0.0571	0.0042	0.0000	0.0014
14	0.2941	0.1235	0.1478	0.0000	0.5340	0.0136	0.6856	0.1010	0.6779	0.5958	0.5911	0.1571	0.0014	0.0000

Corresponding relative frequencies are the following

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	0.0000	0.7728	0.7365	0.0118	0.0773	0.0844	0.2433	0.0999	0.1110	0.4233	0.3260	0.0227	0.0162	0.2970
2	0.7728	0.0000	0.9484	0.0038	0.0059	0.0537	0.0853	0.0787	0.0323	0.2659	0.1688	0.0304	0.0069	0.1300
3	0.7365	0.9484	0.0000	0.0005	0.0267	0.0297	0.1049	0.0577	0.0661	0.2587	0.1606	0.0554	0.0112	0.1534
4	0.0118	0.0038	0.0005	0.0000	0.0019	0.1178	0.0112	0.0703	0.0003	0.0000	0.0000	0.0635	0.2816	0.0000
5	0.0773	0.0059	0.0267	0.0019	0.0000	0.2922	0.5634	0.3135	0.7119	0.4304	0.3662	0.5100	0.0008	0.5545
6	0.0844	0.0537	0.0297	0.1178	0.2922	0.0000	0.1663	0.6885	0.1544	0.2080	0.1508	0.5658	0.0000	0.0152
7	0.2433	0.0853	0.1049	0.0112	0.5634	0.1663	0.0000	0.2872	0.4834	0.7945	0.7833	0.1381	0.0012	0.7057
8	0.0999	0.0787	0.0577	0.0703	0.3135	0.6885	0.2872	0.0000	0.2739	0.1728	0.2128	0.2735	0.0000	0.1080
9	0.1110	0.0323	0.0661	0.0003	0.7119	0.1544	0.4834	0.2739	0.0000	0.3073	0.3003	0.3835	0.0027	0.6922
10	0.4233	0.2659	0.2587	0.0000	0.4304	0.2080	0.7945	0.1728	0.3073	0.0000	0.8231	0.1262	0.0068	0.6103
11	0.3260	0.1688	0.1606	0.0000	0.3662	0.1508	0.7833	0.2128	0.3003	0.8231	0.0000	0.0294	0.0521	0.6070
12	0.0227	0.0304	0.0554	0.0635	0.5100	0.5658	0.1381	0.2735	0.3835	0.1262	0.0294	0.0000	0.0015	0.1652
13	0.0162	0.0069	0.0112	0.2816	0.0008	0.0000	0.0012	0.0000	0.0027	0.0068	0.0521	0.0015	0.0000	0.0005
14	0.2970	0.1300	0.1534	0.0000	0.5545	0.0152	0.7057	0.1080	0.6922	0.6103	0.6070	0.1652	0.0005	0.0000

One may ask at this point whether any simpler stochastic model would be able to generate the same frequencies. In the above model the participants are intrinsically correlated,to forming

the groups the whole matrix A is involved. The following model comes from this idea. Let us offer to the participants of a party to choose *independently* from finitely many possibilities, like:

– to have same food,

– to play hide and seek,

– to watch TV,

– to discuss about Shakespeare,

– to make a small excursion.

Having the program at hand, people have their preferences, they make their choices independently and groups are formed in a natural way. Let us denote by $w(k,s)$ the probability that the k-th participants chooses the s-th possibility than

$$P(\pi(i)=\pi(j)) \quad = \quad \sum_{s=1}^{S} w(i,s) * w(j,s),$$

where S denotes the number of possibilities. If the number of participants goes to infinity the size of groups is the most important feature of the distribution. It may remain bounded or slowly increasing as it is the case in politics when the groups are the political parties having the tendency to become of small number mostly because preference choice. The second possibility is the square root law: the size of groups and their number both are around the square root of c. Third possibility is represented in chemistry: the size of groups remains small and the number of groups increases with c for example for proteins. In the independent model the situation is easily controlled by S but in the pair-potential model we do not know the answer. Our guess is the first possibility on the argument that $Q(\pi, A)$ may achieve the size c for π with bounded elements. In statistical investigations, in cluster analysis partitions appear mostly in the following two different aspects:

– we may form groups from the investigated objects on the basis that any connection is possible only inside the groups ([4], [6])

– we may form the groups of similar objects ([2], citeBolla2).

The second possibility is used in Szemerédi's regularity lemma. Its extension to the pair-potential model is a numbering $f(k),\ k=1,\ldots,c$ of the participants such that

– $1 \le f(k) \le g; k=1,\ldots,c$

– for all $1 \le j \le g$ there is a $1 \le i \le c$ such that $f(i)=j$

– there is a $g*g$ matrix B with elements $b_{u,v}$ such that $a_{i,j}=b_{f(i),f(j)}$ for all $1 \le i < j \le c$

this is called blown-up of the matrix B into matrix A. We investigated partition-clustering in ([13]) and interactive networks in ([14]).

References

[1] O. Barndorff-Nielsen, (1978), *Information and Exponential Families in Statistical Theory,* Chichester: Wiley.

[2] M. Bolla, (2004), Distribution of the eigenvalues of random block-matrices, Linear Algebra and its Applications, **377**, 219-240.

[3] M. Bolla, (2005) Recognizing linear structure in noisy matrices, *Linear Algebra and its Applications*, **402**, pp. 228-244.

[4] M. Bolla and G. Tusnády, (1994), Spectra and optimal partitions of weighted graphs, *Discrete Mathematics*, **128**, pp. 1–20.

[5] S. P. Borgatti and M. G. Everett, (1997), Network analysis of 2-mode data, *Social Networks*, **19**, pp. 243–269.

[6] Fan R. K. Chung, (1997), *Spectral Graph Theory*, CBMS Regional Conference Series in Mathematics, **92**, University of Pennsylvania, Philadelphia, PA.

[7] Davis, Gardner and Gardner, (1941), *Deep South: A Social Anthropological Study of Caste and Class* , University of Chicago Press

[8] P. Frankl and V. Rödl, (2002), Extremal problems on set systems,*Random Structures and Algorithms*, **20,** pp 131-164.

[9] W. T. Gowers,(2006), Hypergraph regularity and Szemerédi's theorem, preprint.

[10] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller, (1953) Equations of State Calculations by Fast Computing Machines, *Journal of Chemical Physics*, **21**, pp 1087–1092.

[11] M. Newman, A-L. Barabási and D. J. Watts, (2006), *The Structure and Dynamics of Networks,* Princeton University Press.

[12] G. Parisi, (1988), *Spin glass theory and beyond*, Word Scientific, Singapore.

[13] L. Rejtő and G. Tusnády, (2005), Clustering methods in microarrays, *Periodica Mathematica Hungarica*, **50**, pp 199–221.

[14] L. Rejtő and G. Tusnády, (2006), Reconstruction of Kauffman networks applying trees, *Linear Algebra and Application*, **417**, pp 220-244.

[15] E. Szemerédi, (1976) *Regular partitions of graphs,* Colloques Internationaux C.N.R.S. **260**, Problèmes Combinatoires et Théories des Graphes, Orsay, pp 399–401.

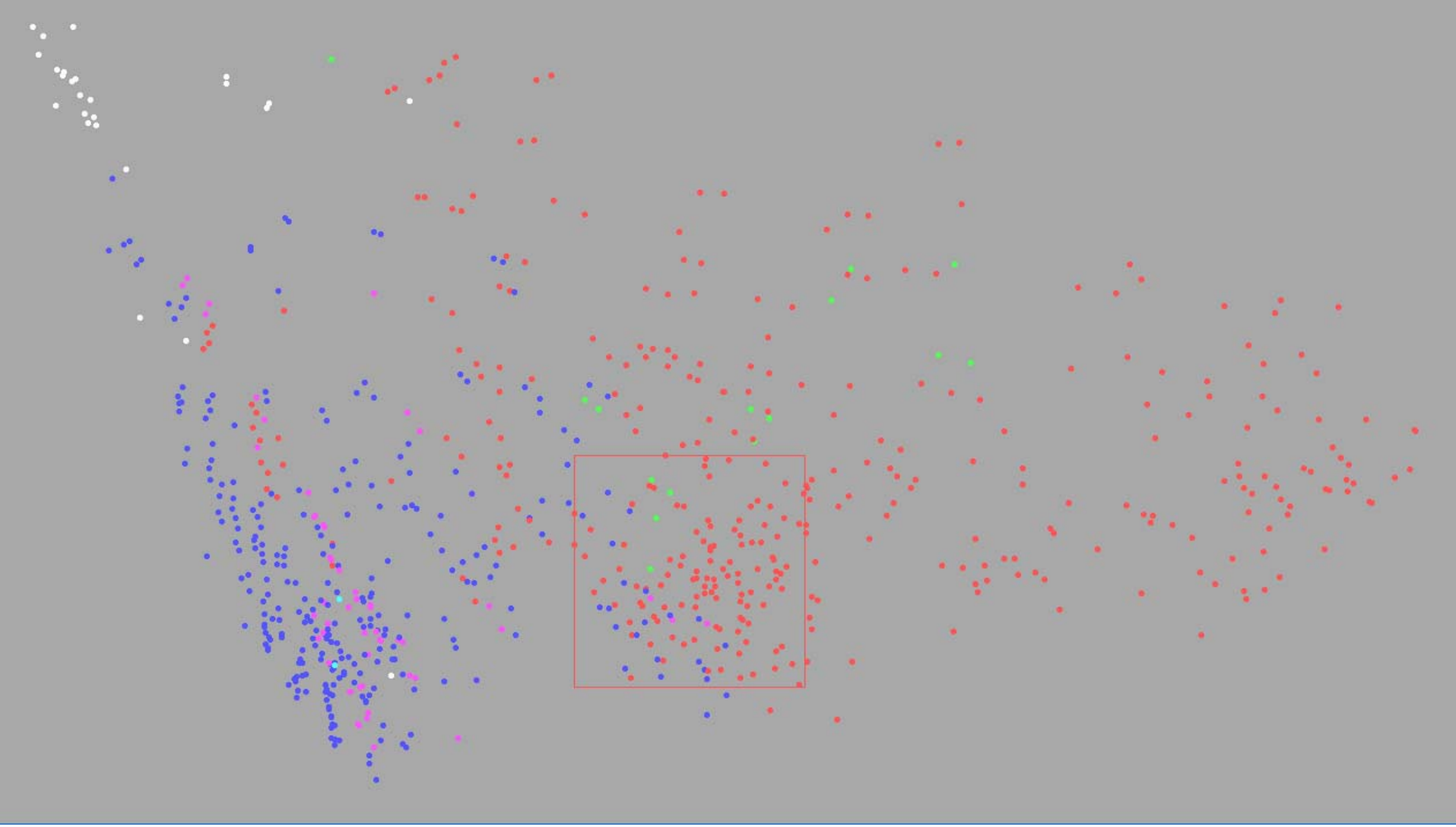

FIGURE 1

Partitions with potential larger than 35 colored after leading members and positioned by contractive algorithm. Partitions in the red rectangle are shown in Figure 3.

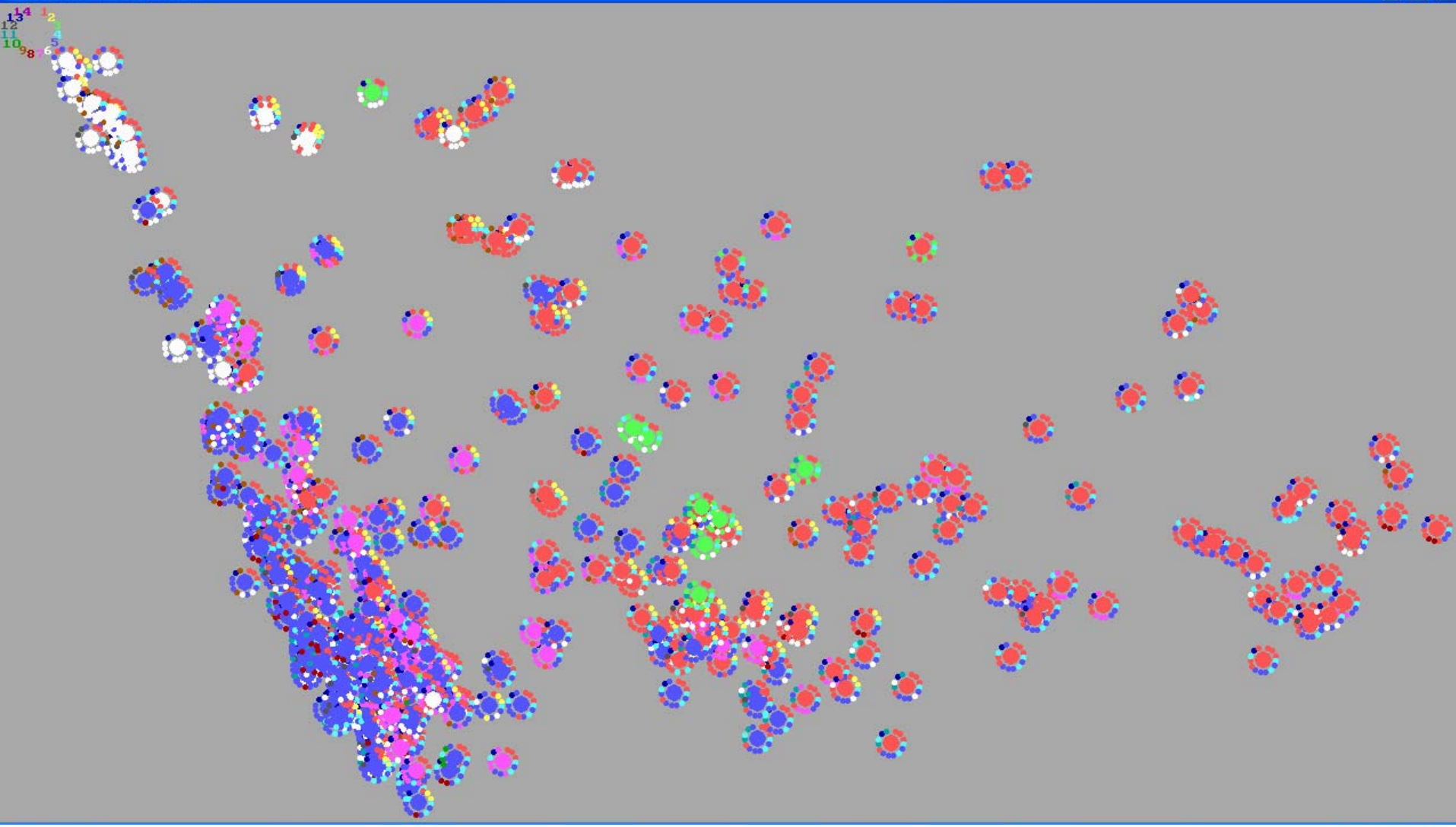

FIGURE 2

Partitions with duration larger than 20,000,000 seconds. Color code of the participants is given in the left upper corner. The color of the center of flowers follows the majority group's color.

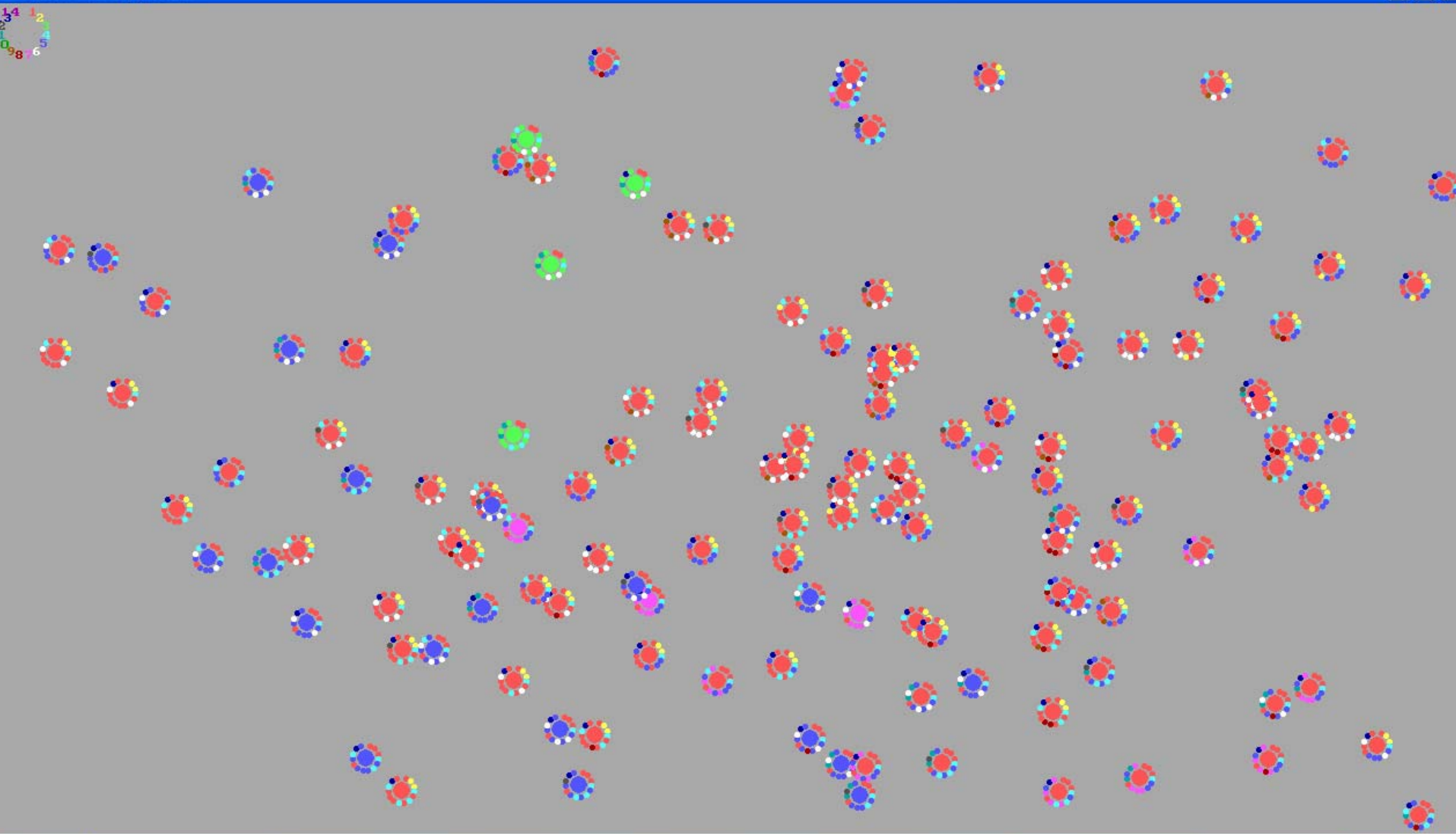

FIGURE 3

Partitions of the red rectangle in Figure 1 with duration larger than 50,000 seconds. Color code of the participants is given in the left upper corner. The color of the center of flowers follows the majority group's color.

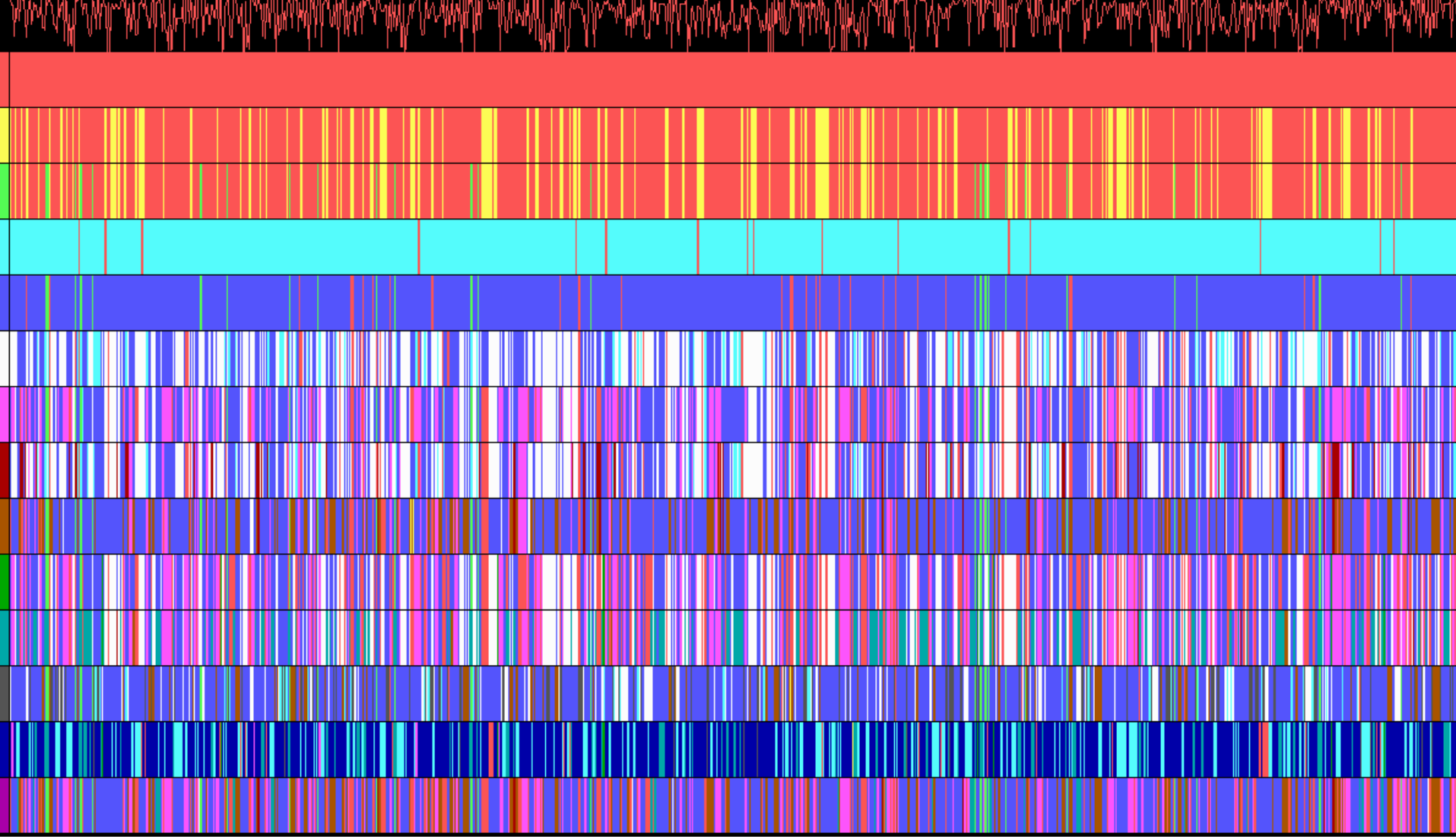

FIGURE 4

Metropolis algorithm with uniting two groups and cutting a group.
Each horizontal line represents one partition with its color coding.
The top red curve shows the change of the potential.

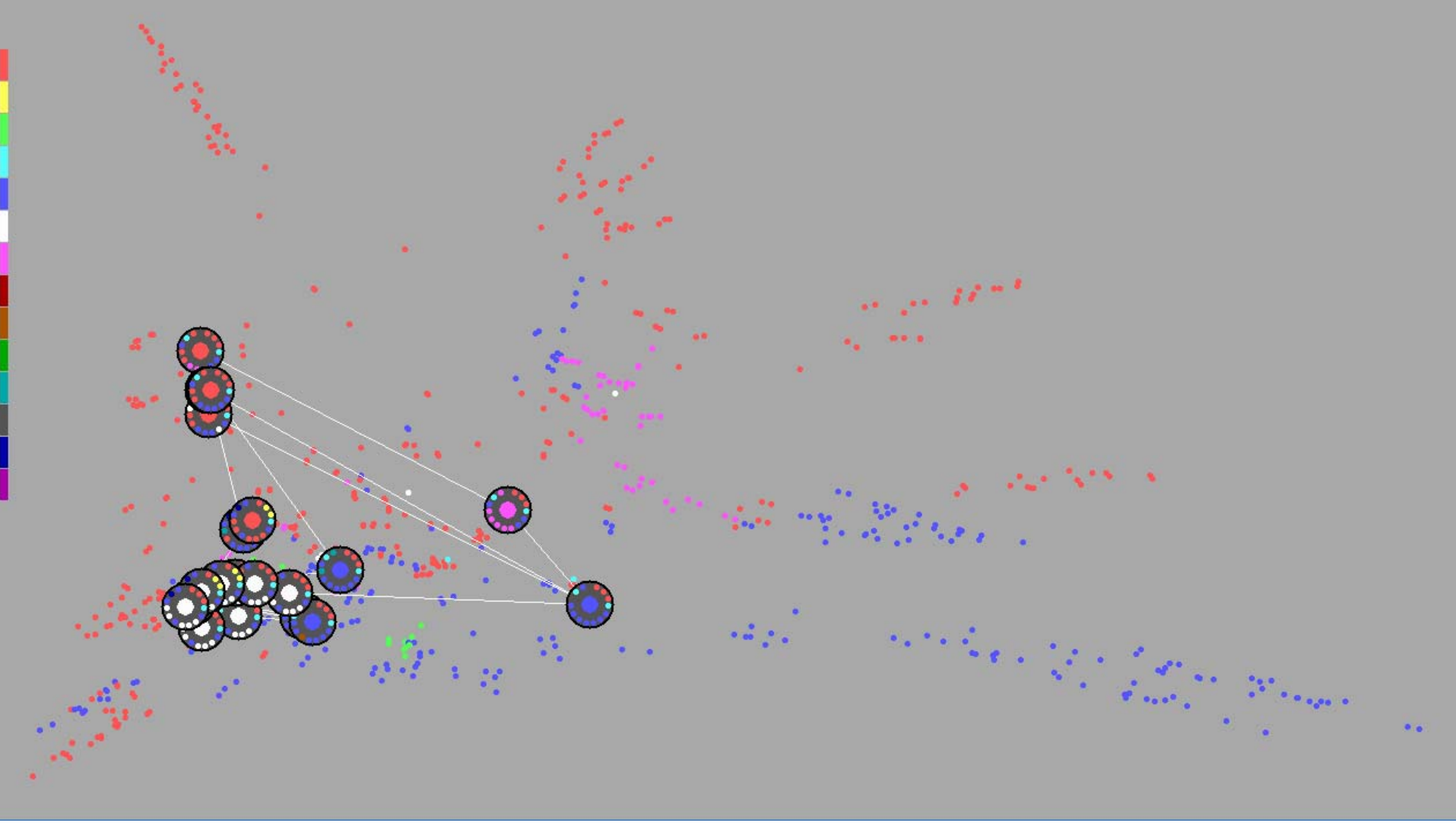

FIGURE 5

Metropolis algorithm with uniting two groups and cutting a group.

The position of the flowers is given by multidimensional scaling. The coloring is the same as in Figures 1-3.

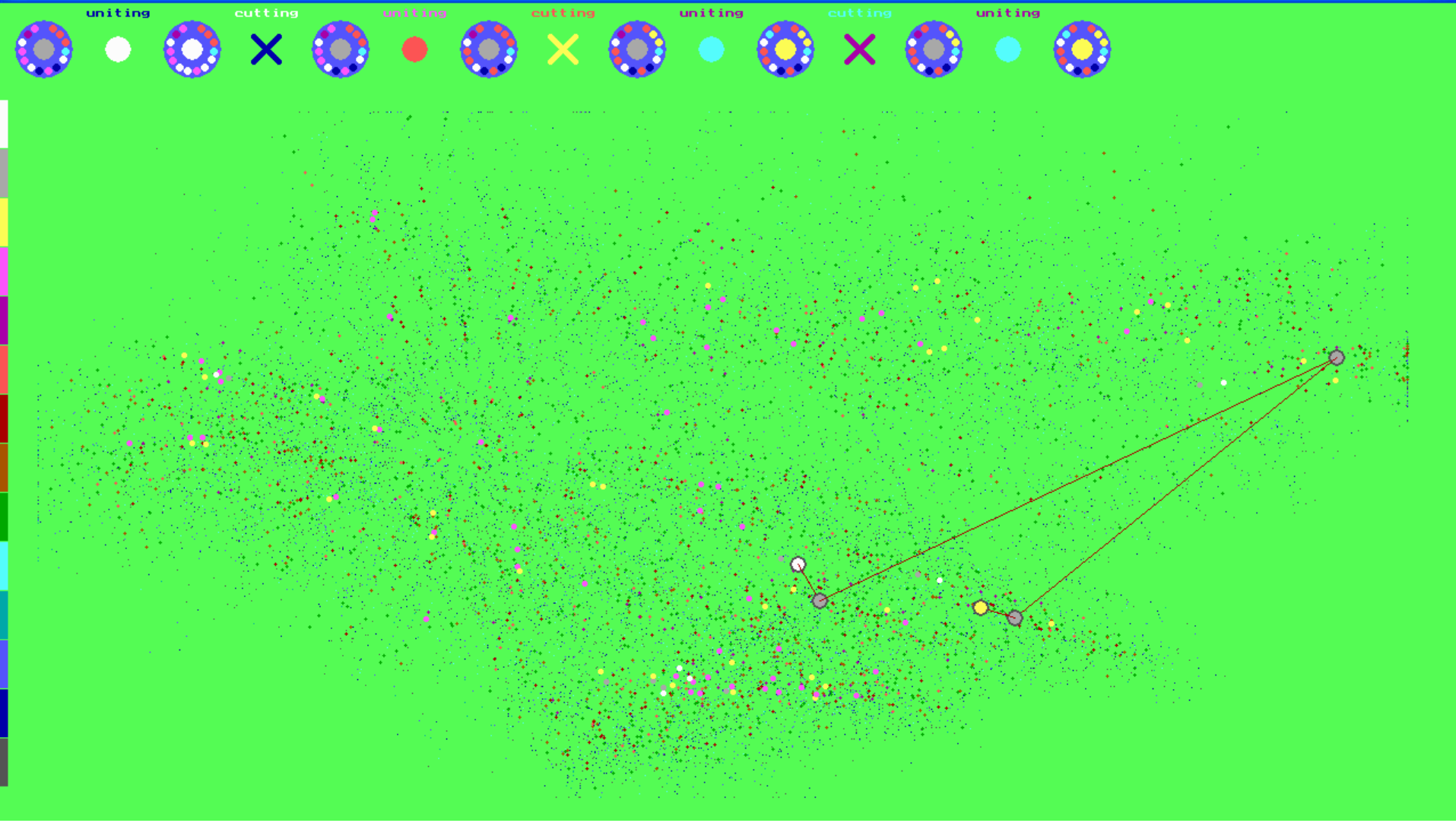

FIGURE 6

Metropolis algorithm with uniting two groups and cutting a group.

In the upper part the colors of the petals show partitions while the color of the center represents the potential of the partition. Color code of the potentials is given on the left of the figure - white means the largest, black the smallest value. All partitions with potentials larger then 30 are mapped using the contracting algorithm.

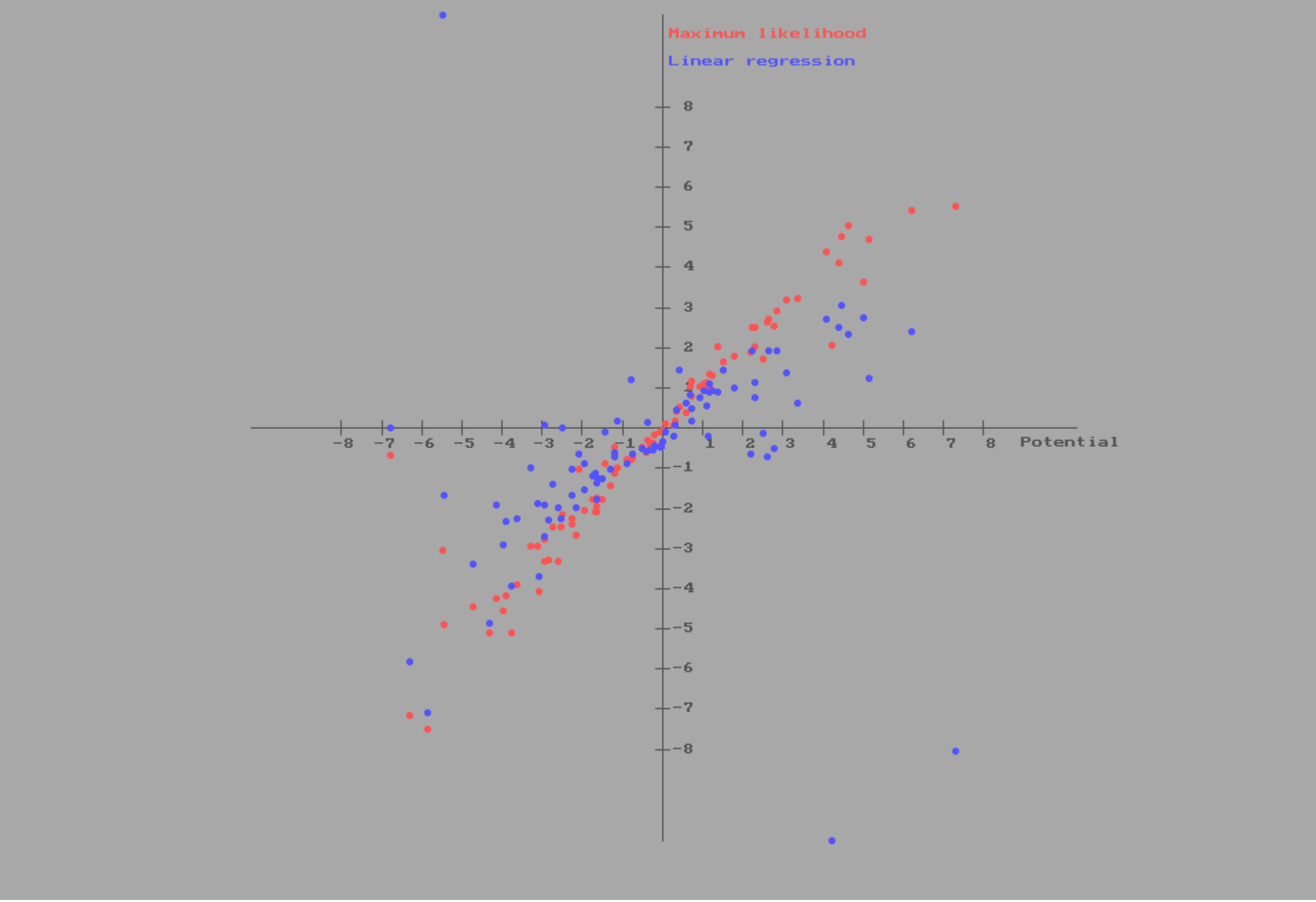

FIGURE 7

Estimation of the potential