# GeneDIVER Visualization and Tools

The HMA hierarchy created by HIMAG shows the different clusters at various densities. The y-axis of the HMA hierarchy represents the different points, and the x-axis represents the different shaving rates. At each resolution, the colored areas of the HMA hierarchy represent the clusters and the black areas represent the pruned background points that do not cluster.

## 1 SIM-2

The Sim-2 dataset used in [1] is a synthetic 2-D dataset sampled from 5 spherical Gaussians and an additional background distribution (Figure 1 (a) & (d)). By using a simple transformation (described in Section 3.4), we get the synthetic Sim-2 Graph dataset (Figure 1 (b)). This dataset not only allows us to more easily test and validate the HIMAG algorithm, but it is also useful for illustrative purposes.

On the HMA hierarchy produced by HIMAG for the Sim-2 Graph dataset (Figure 1 (c)), the small teal cluster at the bottom corresponds to the magenta cluster in the spatial labels (Figure 1 (d)). Its larger sister cluster, which is a slightly lighter teal, is the parent cluster of the cyan, dark blue, red, and green label clusters, which are represented in the HMA hierarchy as the brown and green clusters that form from the large teal parent cluster. Gene DIVER shows the high cluster precision of each cluster, which can be viewed by clicking on the clusters in the hierarchy. This hierarchy illustrates the various diverse and pure clusters visible at different resolutions.

## 2 MARKETING

The marketing dataset we have created consists of two segments: Wellness and Education. The nodes in the graph represent Instagram posts, and the similarity between two posts is measured by the number of shared hashtags.

A good example of the topological relationships between the clusters, is depicted Figure 2 (c). The parent cluster (outlined in white) contains 178 posts about the Law School Admissions Test (LSAT). The bottom gold child cluster contains a subset of 12 posts from the parent cluster about practice LSAT questions. The light purple cluster directly above the gold cluster contains a subset of 27 posts from the parent cluster with the hashtag "#lsatprep". The dark purple cluster directly above the light purple cluster contains a subset of 33 posts from the parent cluster with the hashtag "#lsatstudying". The light brown child cluster on the top was not able to be analyzed because the posts have since been removed. These topological relationships cannot be found with traditional graph partitioning algorithms.

Running the HIMAG algorithm on the Marketing Wellness dataset produced pure clusters of various sizes. Images e–h of Figure 2 illustrate the different classifications of marketing clusters in the Wellness segment. Some clusters are classified as individual influencers—for example, one cluster of this kind contains posts from an individual marketing their weight-loss tea supplement (example post shown in Figure 2 (f)). This cluster has 15 points, and a cluster precision of 100%. Other clusters are classified as professional influencers, such as one cluster of infographic posts from a diet and workout coach (example post shown in Figure 2 (g)). This cluster has 60 points, and a purity of 100%. Lastly, some clusters can be classified as large communities with posts from many different users, like a cluster of posts of weight-loss tips in Portuguese (example post shown in Figure 2 (h)). This cluster has 46 points, and a cluster precision of 100%.

The Instagram URL of each post is embedded into its point on the dataset. Clicking the data point in Gene DIVER opens the Instagram post in the browser, allowing for fast analysis.

## 3 SOCIAL GRAPHS

We present results on two standard social networks datasets; Pokec [3], which is Slovakia's largest social network, and on the LiveJournal social network [2]. These two datasets provide us with two independent of signals that helps us use benchmark with measurements without the need for human labeling. The first signal is the original social graph itself, which is an unweighted graph of connections between people on the social network. The second signal (interests graph) is another independent undirected weighted graph that we build by connecting two people who share "interests" in common based on shared hobbies or community group membership. We use the community profiles that these two datasets provide, namely "hobbies" and "communities" for Pokec and LiveJournal respectively to compute this interests graph.

Then, we set up a learning/prediction problem by splitting the social graph into a training and measurement set; we provide the graph algorithms being compared the training set of the social graph of a random set of edges from the full social graph, while we hold the rest of the graph edges as blind and only use it for measurements to see how well the algorithms predict the missing connections. Additionally, we use the second signal (interests similarity) to weigh the training set graph.

The results are predictions that tend to recommend the friends of friends who have very similar interests at a really high precision and recall, once adjusted for the fact that only a fraction of the good recommendations would already be connected for this highly unsupervised setting; e.g. at 15% precision and 40% recall for Pokec dataset, we are likely recommending all of the close friends whom one should connect with who also share lot of hobbies in common with a given user.

The interests of each node is embedded into the dataset. To view the interests, a user can click on the data point in Gene DIVER to open the hobbies in a Google search. The hobbies for Pokec are inspectable (since they are names of hobbies), whereas the communities for LiveJournal are not inspectable (since they are community IDs). Therefore, in Figure 3, we are only presenting results from Pokec.

## REFERENCES

[1] Gunjan Gupta, Alex Liu, and Joydeep Ghosh. 2008. Automated Hierarchical Density Shaving: A robust, automated clustering and visualization framework for large biological datasets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 4, 7 (2008).

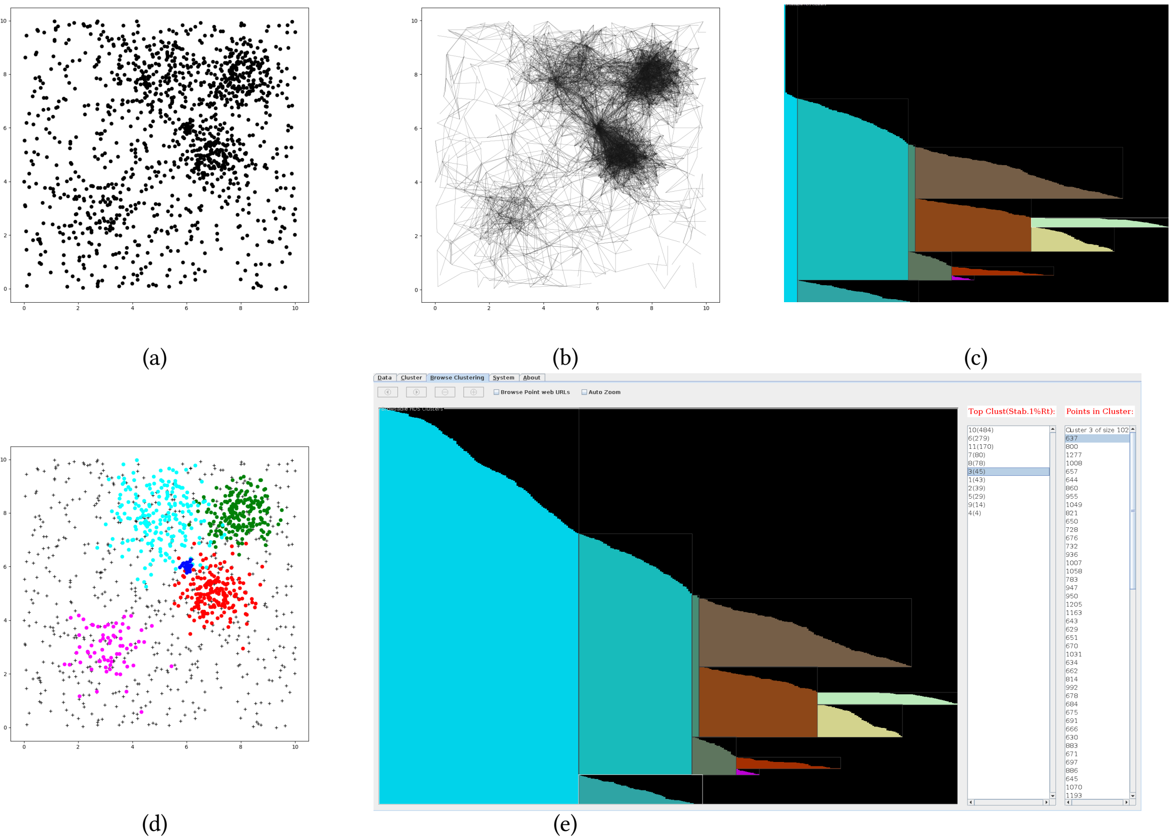


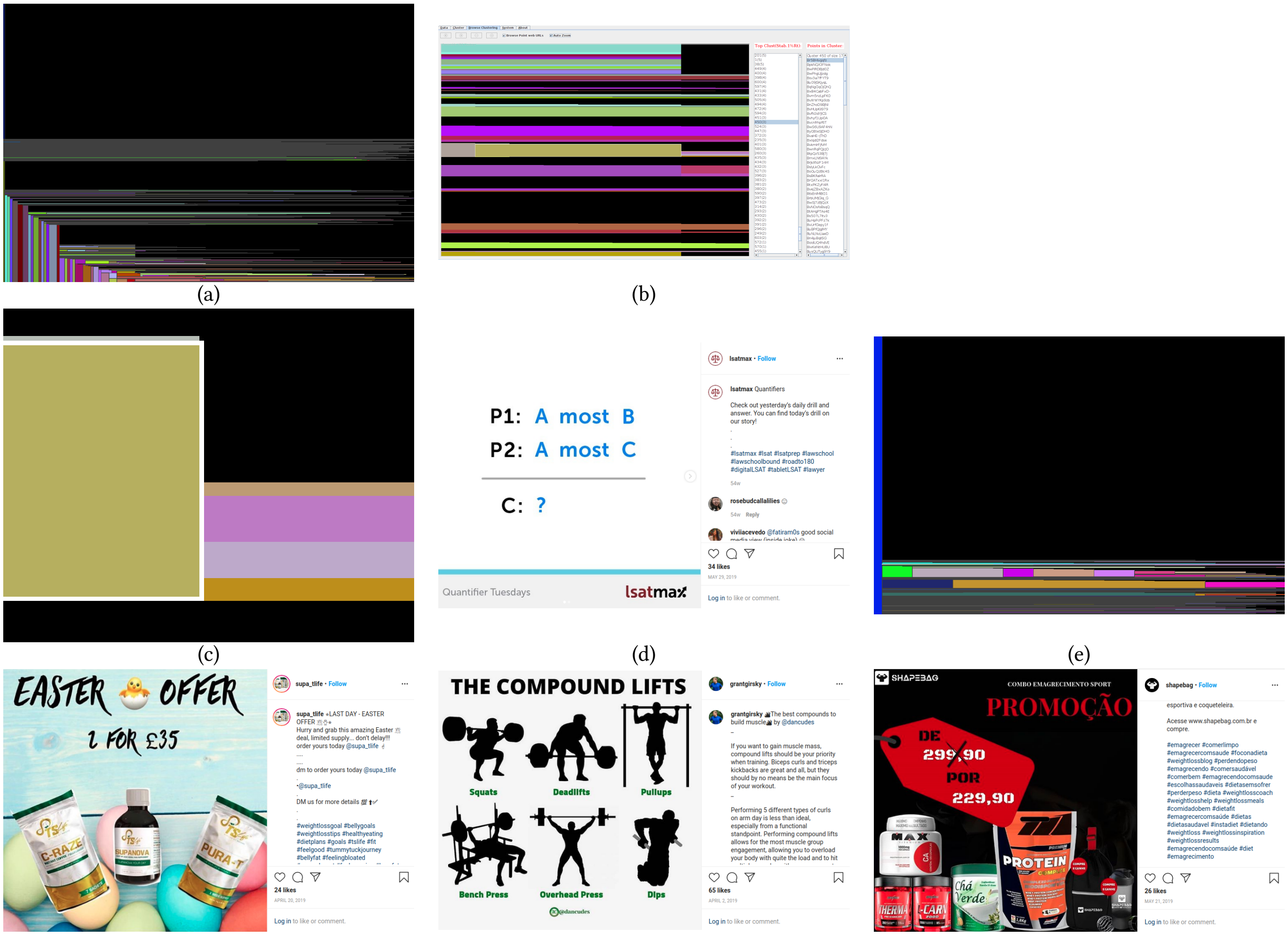
**Figure 1: Sim-2 synthetic dataset. (a) Unlabeled points in 2-D euclidean space as seen by spatial clustering algorithm; (b) a sample of edges with $s_\epsilon \geq 0.9$ from the Sim-2 Graph data; (c) HIMAG results with $s_\epsilon = 400$ shows visually identical topology and clusters to spatial clustering; (d) labeled 2-D data points used for measurements showing 5 labeled *dense* clusters in color, and the background (noise) points in black; (e) full Gene DIVER view of HMA hierarchy produced by HIMAG.**

**Table 1: Social experiments graph sizes and parameters.**

| **Social Network** | **Pokec** | | **LiveJournal** | |
|---|---|---|---|---|
| **No. of interests** | 76,914 hobbies | | 287,512 communities | |
| **Top Interests** | **Count** | **Freq. Thresh.** | **Count** | **Freq. Thresh.** |
| **Top Interests** | 12,285 | 2 | 1,489 | 500 |
| **Graph sizes** | **Node count** | **Edge count** | **Node count** | **Edge count** |
| **Social Graph** | 1,632,803 | 30,622,564 | 3,997,962 | 34,681,189 |
| **Top Interests Graph** | 758,054 | 10 billion+ | 612,577 | 20 billion+ |
| **Intersected Graph** | 736,120 | 22,411,849 | 612,577 | 10,070,149 |
| **Training Graph** | 208,620 | 151,336 | 212,076 | 201,158 |
| **Measurement Graph** | 711,950 | 7,420,344 | 612,007 | 9,868,991 |
| **Training Social Graph Fraction** | 2% | | 2% | |

[2] A. Dasgupta M. Mahoney. J. Leskovec, K. Lang. 2009. Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters. *Internet Mathematics* 6, 1 (2009), 29–123.

[3] Lubos Takac and Michal Zabovsky. 2012. Data analysis in public social networks. *International Scientific Conference and International Workshop Present Day Trends of Innovations* (May 2012), 1–6.



lsatmax • Follow

lsatmax Quantifiers

Check out yesterday's daily drill and answer. You can find today's drill on our story!
.
.
.
#lsatmax #lsat #lsatprep #lawschool #lawschoolbound #roadto180 #digitalLSAT #tabletLSAT #lawyer

54w

rosebudcallalilies 😑

54w Reply

viviiacevedo @fatiram0s good social media view (inside joke)

34 likes

MAY 29, 2019

Log in to like or comment.

supa_tlife • Follow

supa_tlife ⭐LAST DAY - EASTER OFFER ⭐

Hurry and grab this amazing Easter deal, limited supply... don't delay!!! order yours today @supa_tlife
....
....
dm to order yours today @supa_tlife
.
•@supa_tlife
.
DM us for more details 💯⬆✔
.
.
#weightlossgoal #bellygoals #weightlosstips #healthyeating #dietplans #goals #tslife #fit #feelgood #tummytuckjourney #bellyfat #feelingbloated

24 likes

APRIL 20, 2019

Log in to like or comment.

grantgirsky • Follow

grantgirsky The best compounds to build muscle by @dancudes
–
If you want to gain muscle mass, compound lifts should be your priority when training. Biceps curls and triceps kickbacks are great and all, but they should by no means be the main focus of your workout.
–
Performing 5 different types of curls on arm day is less than ideal, especially from a functional standpoint. Performing compound lifts allows for the most muscle group engagement, allowing you to overload your body with quite the load and to hit

65 likes

APRIL 2, 2019

Log in to like or comment.

shapebag • Follow

esportiva e coqueteleira.

Acesse www.shapebag.com.br e compre.

#emagrecer #comerlimpo #emagrecercomsaude #foconadieta #weightlossblog #perdendopeso #emagrecendo #comersaudável #comerbem #emagrecendocomsaude #escolhassaudaveis #dietasemsofrer #perderpeso #dieta #weightlosscoach #weightlosshelp #weightlossmeals #comidadobem #dietafit #emagrecercomsaúde #dietas #dietasaudavel #instadiet #dietando #weightloss #weightlossinspiration #weightlossresults #emagrecendocomsaúde #diet #emagrecimento

26 likes

MAY 21, 2019

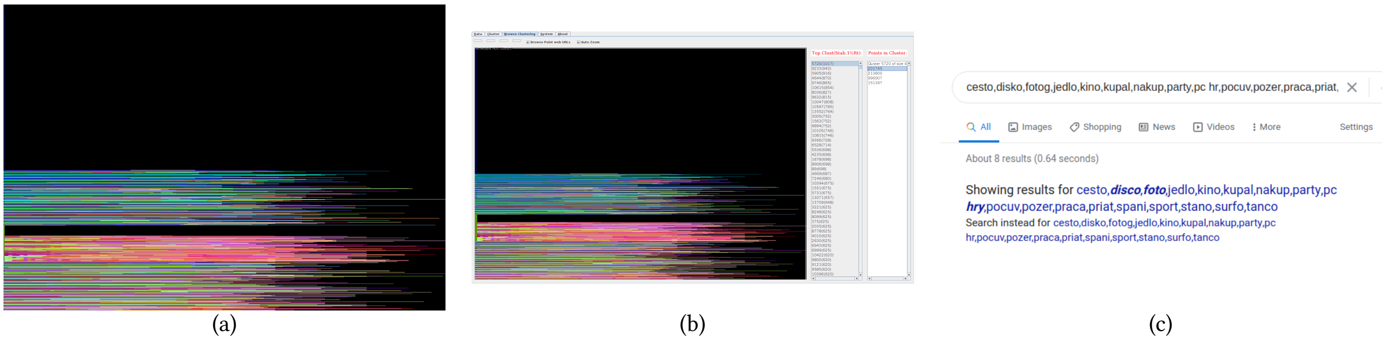
Log in to like or comment.

(a) (b)

(c) (d) (e)

(f) (g) (h)

**Figure 2: Marketing Education dataset: (a) full HMA hierarchy produced by HIMAG for the Marketing Education dataset; (b) zoomed in full Gene DIVER view of HMA hierarchy produced by HIMAG; (c) example of the LSAT topology in HMA hierarchy; (d) post that is part of the "LSAT practice question" child cluster of the LSAT cluster. Marketing Wellness dataset: (e) full HMA hierarchy produced by HIMAG for the Marketing Wellness dataset; (f) advertisement for an Easter sale for a weight-loss supplement; (g) infographic by a diet and workout coach explaining the best types of compound lifts to build muscle; (h) Portuguese promotion of a weight-loss powder.**



cesto,disko,fotog,jedlo,kino,kupal,nakup,party,pc hr,pocuv,pozer,praca,priat,

About 8 results (0.64 seconds)

Showing results for cesto,***disco***,***foto***,jedlo,kino,kupal,nakup,party,pc ***hry***,pocuv,pozer,praca,priat,spani,sport,stano,surfo,tanco

Search instead for cesto,disko,fotog,jedlo,kino,kupal,nakup,party,pc hr,pocuv,pozer,praca,priat,spani,sport,stano,surfo,tanco

(a) (b) (c)

**Figure 3: Pokec (social network) dataset: (a) full HMA hierarchy produced by HIMAG; (b) zoomed in full Gene DIVER view of HMA hierarchy produced by HIMAG; (c) example hobbies of a user shown as a Google search.**