CS224W Project Report: Finding Top UI/UX Design Talent on Adobe Behance

Susanne Halstead, Daniel Serrano, Scott Proctor

6 December 2014

1 Abstract

The Behance social network allows professionals of diverse artistic disciplines to exhibit their work and connect among each other. We investigate the network properties of the UX/UI designer subgraph. Considering the subgraph is motivated by the idea that professionals in the same discipline are more likely to give a realistic assessment of a colleagues work. We therefore developed a metric to assess the influence and importance of a specific member of the community based on structural properties of the subgraph and other measures of popularity. Furthermore, we investigated spacial properties of the social graph. We identified appreciations as a useful measure to include in a weighted PageRank algorithm, as it adds a measure of prestige to the ranking that is not contained in the structural information of the graph. With this pagerank, we identified locations that have a high density of influential UX/UI designers. Investigating the geographic distribution of the Behance social graph further, we found that there are two separate processes that generate connections, a geographically dense circle of presumably persons from the direct professional environment of the artist and a geography independent network of connections.

2 Introduction

Professionals of all disciplines form social networks of relationships from attending the same schools, being coworkers, participating in the same local job market, attending the same conferences, and active participation in professional communities. For some professions, online communities have emerged. We are examining the community of User Experience (UX)/ User Interface (UI) designers present in Adobe's Behance social network. Behance allows creative professionals of various artistic background to share their work, receive feedback and network with each other. Importance and prestige of artists in the community is currently calculated on the whole network and based on raw counts of followers and appreciations. We take a different approach, segmenting the network into areas of practice and using graph techniques to calculate importance.

For that, we are adapting algorithms for finding influencers in social networks to reflect the existence of several measures of influence: number of followers, and number of appreciations, number of views. Additional information available is design schools attended and current geographic location.

2.1 Problem Description

In the 'People' perspective, Behance currently identifies artists' creative fields, and ranks users on either views, or appreciations, or most followers. From our assessment, these rankings are purely based on ordering search results on raw counts of these measures, and these measures are calculated over the whole network, not over the network induced by applying a filter on creative field. We want to assess if segmenting the social graph by artistic field and applying a topology aware algorithm weighted with the artists' prestige to calculate influence of a node as well as considerations of locality will add insights. We want to address the following topics for the community of UX/UI designers:

1. Find influential UX/UI designers based on their position in the network and their prestige.
2. Test different known strategies for influence maximization on the network.
3. Determine significant geographic clusters and find influencers in the specific geographies.
4. Analyze whether the geographic distribution of artists linked on Behance suggest a social network that is based on personal connections or if connections are not strongly linked to geography, suggesting a virtual community rather than a community created of people that know each other outside of Behance.

These results can have immediate practical use: A limitation to specific artistic field for the calculation of important nodes lays the groundwork for more focused marketing campaigns or hiring. The determination of significant geographic clusters of professionals supports decisions of where to conduct user group meetings, conferences or where to physically place design shops to be able to hire top talent. An assessment of how indicative a connection in Behance is of a connection on a personal level gives insight what type of influence exists between them and thus what type of messages might be most effective.

3 Related Work

3.1 On Ranking and Influence

In an effort to discover the content of the web, there are several web mining approaches that play together, namely web content mining (Information Retrieval techniques), web structure mining and web usage mining [9]. Web content mining is used to group web pages by subject matter, while web structure mining approaches have been developed to determine a ranking by importance of webpages. There are two important algorithms in web structure mining: PageRank and HITs. While originally developed for web search engine applications, the algorithms are applicable to ranking nodes in other types of networks. Bento [2] illustrates how these algorithms have relevance to the task of finding important nodes (influencers) in social networks, showing that the concepts used for web structure mining reasonably translate to finding important nodes in social graphs.

Rankings derived from graph structure mining directly translate to the notion of influence. Y. Singer [8]; addresses the topic of how to best select a subset of influencers to maximize (commercial) message propagation given resource constraints of a campaign, such as paying influencers to post commercial messages, writing blog posts, or giving out free trials of new products. The global optimization problem is known to be NP-hard. There has been a number of proposed -approximations with better runtime characteristics. Singers work in particular considers the aspect, that not all influencers have the same cost of becoming early adopters.

3.1.1 The PageRank Algorithm

PageRank [6] is an algorithm developed for web structure mining in support of ranking search result by relevancy. In a first step a set of pages that match a given search query by keywords is retrieved. Then PageRank for this set is calculated to put the results in order of priority. PageRank assumes that if a page has important links to it, its linked pages are also important; it therefore takes back links into account. A page's ranking is high if the sum of the ranks of its back links are high. This concept is expressed in the simplified PageRank formula:

$$PR(u) = c \sum_{v \epsilon B(u)} \frac{PR(v)}{N_v}$$

With u representing the webpage; $B(u)$ is the set of webpages that link to u. $PR(u)$ and $PR(v)$ are the PageRank scores of u and v; c is a normalization factor. The PageRank of a page is evenly distributed between its outgoing links. PageRank is calculated iteratively, until convergence is reached.

A modification of the basic formula solves the 'rank sink' problem (accumulation of rank in loops of pages with no outlink). PageRank is extended with a 'teleport' function, expressed as a dampening factor in the page rank formula. The dampening factor d is the probability of a random surfer following a link on the page; $1 - d$ is the probability to teleport to any page. In web structure mining, the dampening factor is often set to 0.85, however, it is frequently adjusted for other contexts to reflect observed behaviors in the domain.

$$PR(u) = (1 - d) + d \sum_{v \epsilon B(u)} \frac{PR(v)}{N_v}$$

3.1.2 Weighted Ranking Approaches

Xing and Ghorbani [3, 9] point out a shortcoming in the PageRank algorithm, which is that all links are treated with equal weight. Weighted PageRank implementations modify the probability of an outlink being used, and with that the PageRank of the linked pages, by weighing outedges. A general formulation of the weighted PageRank formula used by Xing and Ghorbani is:

$PR(u) = (1 - d) + d \sum_{v \epsilon B(u)} PR(v) W_{(u,v)}$

The definition of W will vary by context. In their paper [9], Xing and Ghorbani extend PageRank to consider the popularity of a webpage, as measured by the number of inlinks and outlinks. They achieve to produce more relevant results using this approach. Once we move from the realm of web structure mining to other types of networks, the definition of what useful weights may change. Y. Ding points out that in the context of citation networks there is a difference between popularity and prestige. In this context, popularity is defined as the number of times a researcher is cited; prestige is defined as the number of times the author is cited in highly cited papers. The finding is that prestige is most important when assigning weights in order to produce most relevant results, and define a measure based on the definition of prestige as the weights.

3.2 Geographic Distance Distributions in Social Networks

Scellato et al. [7] evaluate the socio-spatial properties of various social networks. They show that there is a different relationship between the probability of a link and geographic distance between users, depending on what type of interactions the social network provides and how likely its users are to also have an off-line relationship, however, there also exist robust universal features observed in their samples. They propose several models to describe these effects. Lambiotte et. al inversigate the geography of links in mobile phone communication networks (the social network with the strongest personal ties) [4]. They find that the probability of finding a link of length d follows the distribution $P(d) \propto d^2$. Backstrom et. al find that Facebook shows strong ties between geography and probability of friendship [1] with $P(d) \propto \frac{1}{d}$. Liben-Novell et. al found that bloggers on live journal [5] connect with probabilities $P(d) \propto \frac{1}{d} + \epsilon$, where ϵ is a constant probability which exists regardless of distance. Hence Live Journal acts more like a virtual community, where distance plays much less of a role.

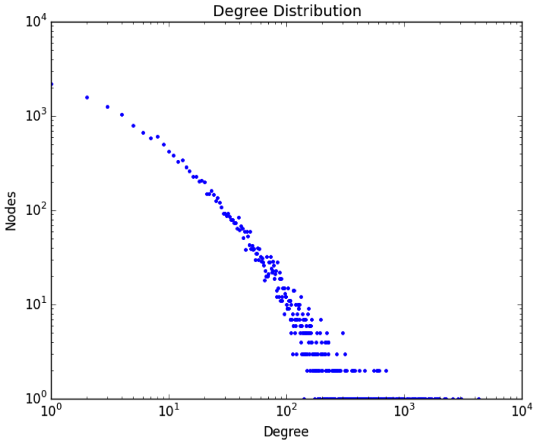
4 Data

The data for our study was obtained via web scraping of the Behance website. We started by collecting profiles of UX/UI designers from the search page. Using these search results as seeds, we then proceeded to scrape the profile information and statistics. The profile information contains the profile id, the list of artistic interests, location of the artist (city, state, country), count of followers, count of following, project views , profile views, project appreciations, project comments. The relationships of followers and following form directed links between nodes. In order to extract the subgraph of UX/UI designer, we followed all outgoing links per node (following). This collects all edges, including edges leading outside the subgraph. The edges leading outside the subgraph were discarded. This process yields the subgraph of UI/UX designers and their connections among each other. In a next step, we geocoded the locations of the profiles with latitudes and longitudes at city center level using the Bing Maps API. Table 1 summarizes some properties of the graph of UX/UI :

The degree distribution of the UX/UI designer graph follows a power law distribution.

Measure	
Number of Nodes	16,885
Number of Edges within UX/UI Network	240,653
Full Diameter of Network	11
Clustering Coefficient	0.168684

Table 1: Properties of the UX/UI Designer Subgraph



5 Finding Important Nodes

In the following, we develop an approach to implementing a weighted page rank for the Behance network.

5.1 Page Rank Ranking

For this step, we used the Snap PageRank function, to append the page rank to each profile in the UX/UI designer subgraph, discarding all nodes that are not connected by at least one link to the graph. In a next step, we investigated whether the PageRank of a profile is correlated to the measures of followers, following, project views, project appreciations. Table 2 summarizes the correlation factor between the various measures:

	PageRank	Followers	Following	ProjectViews	ProfileViews	ProjectAppreciations	ProjectComments
PageRank	1.0000	0.8069	0.1279	0.8017	0.7048	0.7108	0.7280
Followers		1.0000	0.2409	0.7300	0.7426	0.6387	0.6450
Following			1.0000	0.1314	0.3365	0.1392	0.1850
ProjectViews				1.0000	0.7462	0.9168	0.8620
ProfileViews					1.0000	0.7476	0.7565
ProjectAppreciations						1.0000	0.8748
ProjectComments							1.0000

Table 2: Correlation Coefficient between Profile Properties

5.2 Weighted Page Rank Ranking

We observe that page rank and the number of followers are highly correlated. Therefore, including the numbers of followers into the PageRank calculation would not convey new information. After the number of followers, the number of appreciations is an important measure of the prestige of an artist on Behance. The number of appreciations is not as directly correlated to the PageRank score. We therefore extend the basic PageRank formula to contain the relative prestige among the linked nodes as weights for the outlinks at each node. Our modified PageRank formula is:

$$PR(u) = (1-d) + d\sum_{v\epsilon B(u)} \frac{PR(v)A(v)}{TA(u)}$$

With the notations: d dampening factor, u profile page of a Behance user, $B(u)$ is the set of users that follow user u - link to user u on the graph, $PR(u), PR(v$ are the rank scores of pages u and v, $A(v)$ is the number of appreciations for page v, $TA(v)$ is the total number of appreciations of pages pointing to u, thus $TA(v) = \sum_{v\epsilon B(u)} A(v)$.

The Behance social graph of UX/UI designers is small enough to be handled by a single machine. We implemented the basic PageRank formula in its matrix formulation using Python, using numpy sparse matrix algebra.

6 Geographic Distribution

The distance distribution of links in the Behance UX/UI designer graph is a bimodal distribution. This suggests that there are two underlying processes at work by which people connect. A possible interpretation is that the geographically shorter connections in the first mode represent connections between people that personally know each other (from attending the same design school, coworkers), while the second mechanism of connection is following the work of an artist independent of geographic location. We used the Expectation Maximization approach for parameter estimation of overlapping Gaussians implemented in the CRAN R package mixtools.normalmixEM to determine mean and standard deviation of the two distributions.

	Mode 1	Mode 2
Mean in km	983.767329	7982.796455
Standard Deviation in km	803.865690	3670.625458

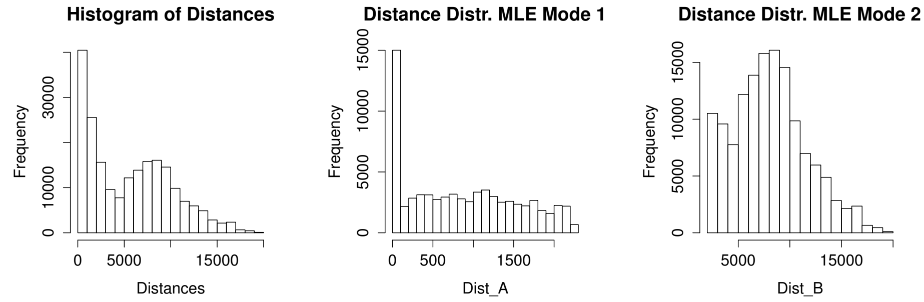


Figure 1: Separating the Data Assuming Overlapping Gaussians

With the two gaussian distributions as defined in table 6, we separated the data by comparing their z-scores under both models. The datapoint was attributed to the model with the lowest z-score. This separation indicates, that there is a part of connections concentrated in close vicinity of users, that is separate from the second process. We interpret that this reflects connections among people that most likely know each other from school or work, and hypothesize that the second process is geography independent. In order to test this idea, we construct a random graph of 30,000 links on the nodes, to generate the distance distribution when links are created independent of geography. This model creates a baseline of what distributions connections would have that were created independent of location. We then again used z-scores to separate the data, this time, attributing all data points that were within two standard deviations of the random model to a geography independent process, and labeling remaining points as geography dependent. This view supports our idea that part of the links are geography dependent and in close vicinity, while another part of links is created independent of geography.

7 Where do the Influencers Live

As a concluding consideration, we investigated what locations in the US have the most influential UX/UI designers. PageRank allows us to have a clearer idea. When only considering the density of users (Figure 3, Tile 1), many cities

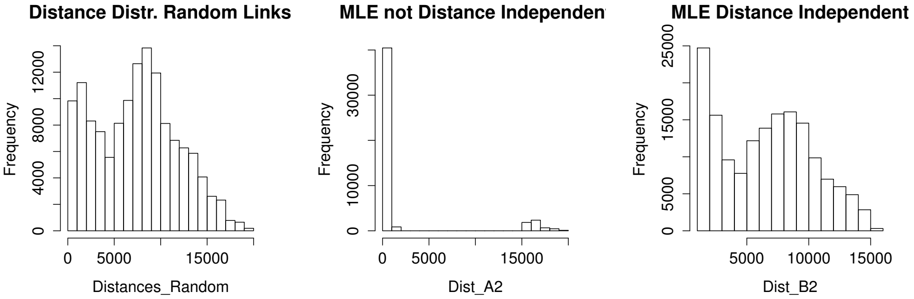


Figure 2: Separating Data by Comparison with Random Model

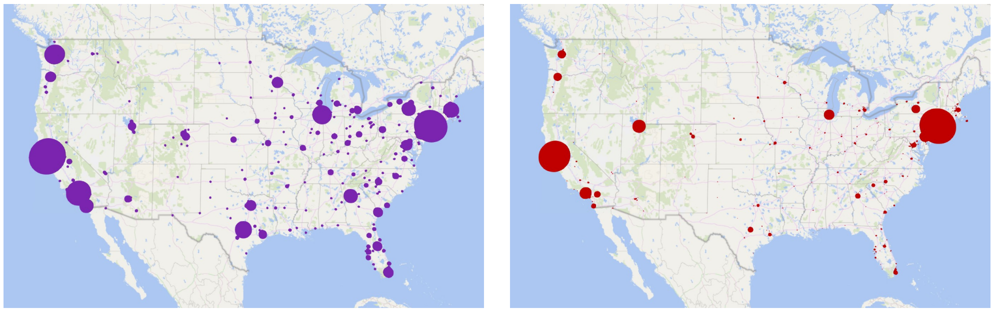


Figure 3: Geographic Distribution of Behance Users: 1 By User Density 2 By Page Rank Density

in the US come out to have a significant number of designers. However, when considering PageRank density (Figure 3, Tile 2) as the measure, New York and San Francisco stand out as the clear epicenters of influence in UX/UI design in the US.

8 Discussion and Future Work

Our work was able to show that PageRank is a useful measure of influence in the Behance social network, and that specifically extending PageRank to contain scores of appreciation allow the identification of the most important UX/UI designers.
Future work would include a more robust evaluation of the relevance of the adjusted PageRank measure.

References

[1] Lars Backstrom, Eric Sun, and Cameron Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th international conference on World wide web*, pages 61–70. ACM, 2010.

[2] Carolina Bento. Finding influencers in social networks.

[3] Ying Ding. Applying weighted pagerank to author citation networks. *Journal of the American Society for Information Science and Technology*, 62(2):236–245, 2011.

[4] Renaud Lambiotte, Vincent D Blondel, Cristobald de Kerchove, Etienne Huens, Christophe Prieur, Zbigniew Smoreda, and Paul Van Dooren. Geographical dispersal of mobile communication networks. *Physica A: Statistical Mechanics and its Applications*, 387(21):5317–5325, 2008.

[5] David Liben-Nowell, Jasmine Novak, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins. Geographic routing in social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(33):11623–11628, 2005.

[6] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. 1999.

[7] Salvatore Scellato, Anastasios Noulas, Renaud Lambiotte, and Cecilia Mascolo. Socio-spatial properties of online location-based social networks. *ICWSM*, 11:329–336, 2011.

[8] Yaron Singer. How to win friends and influence people, truthfully: influence maximization mechanisms for social networks. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 733–742. ACM, 2012.

[9] Wenpu Xing and Ali Ghorbani. Weighted pagerank algorithm. In *Communication Networks and Services Research, 2004. Proceedings. Second Annual Conference on*, pages 305–314. IEEE, 2004.