Examining Homophily in Flickr Friendships

Jennie Kim, Amber Latham, and Jaime Sanchez

California State University, Long Beach

Flickr is an online photo sharing platform that promotes two main goals: to help users make their photos available online and to enable new ways of organizing photos and videos (Flickr, 2018). The platform can be accessed online from both computers and mobile devices, allowing users to seamlessly share and collaborate on photo albums. Currently, Flickr offers paid membership services. As a Flickr member, you can upload, organize, share, and collaborate on photo albums. Non-members may also view your photo albums.

This particular social network is interesting because it explores the connections that are specifically influenced by online content. While most online social networks have overlap between real-life and online connections, friendships on Flickr may have less overlap because it is specifically designed for global image sharing and networking beyond the user's real-life connections. Therefore, there are several hundred groups that have formed on Flickr thus far, ranging from those who share the same taste in camera equipment to those who have the same style or preference of photography. The data provide a unique opportunity to observe patterns of homophily within friendship ties that exist primarily online. This may provide insight on sub-group behaviors of photographers and reflect social network patterns in a user-generated online society. The connections on Flickr provide meaningful data attributing to how online social hubs form.

It is hypothesized that users' friendships ties follow the principle of homophily, wherein users will have the most relationship ties with other users inside their group. Homophily has been extensively studied in social networks. With the increasing prevalence of social media, the study of homophily has evolved to include online friendships as well. Tang and Liu (2009) generated the Flickr dataset used in our analysis in order to study community detection, which is a phenomenon where people interact more frequently within groups than between groups. Faralli, Stilo and Velardi (2015), investigated homophily among Twitter users by clustering them based on interest similarity. Homophily was then calculated as average similarity within the generated cluster.

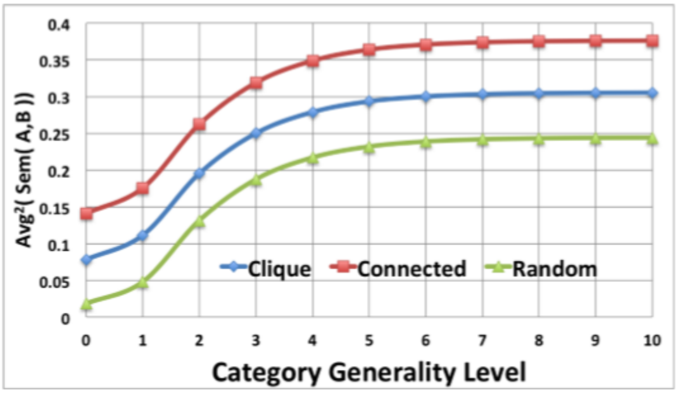


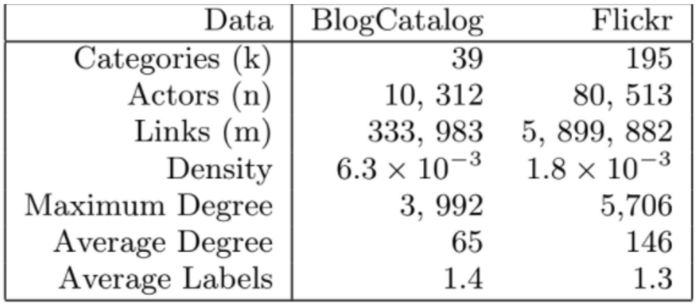
Figure 1. *Faralli, Stilo & Velardi, 2015*

As shown in Figure 1, Twitter users had higher homophily (y-axis) when “connected.” Connected refers to a cluster of users who have friendship ties with each other, “clique” refers to a cluster generated by similar interests between users, and random refers to a cluster with users randomly shuffled among the different clusters generated. Although some clusters were artificially generated and not representative of actual ties, this exemplifies online homophily and how it is stronger within groups than between groups.

Data

The Flickr dataset used for this analysis is a subset derived by Tang and Liu in 2009 of a much larger Flickr dataset, which was accessed and downloaded via Arizona State University’s Social Computing Repository (Tang & Liu, 2009). The file consists of nodes (users), groups (attributes) and edges (friendships). The entirety of the subset consists of 80,513 nodes, 195 groups, and 5,899,882 edges (Tang & Liu, 2009). Attributes are anonymized, with groups identified by numbers rather than names. Additionally, the group membership data is binary (i.e., users are either in a group or not in a group) and does not include level of activity or engagement within each group by each member.

The typical Flickr user base consists of professional or seeking-to-be professional photographers (Dotan & Panayiotis, 2010). Users can be part of multiple groups, subscribing to an average of 1.3 groups (Table 1). Each edge consists of an undirected tie, or mutual relationship, between users. It is important to note that the nature of Flickr friendships has changed since 2009 when the dataset used in this analysis was derived. Whereas friendships on Flickr began as mutual, undirected ties (similar to friendships on social media networking sites such as Facebook), they have since evolved to a directed, following relationship in which a user may follow another user without that user following them back (similar to friendships on social media networking sites such as Twitter and Instagram).



Data	BlogCatalog	Flickr
Categories (k)	39	195
Actors (n)	10, 312	80, 513
Links (m)	333, 983	5, 899, 882
Density	6.3×10^{-3}	1.8×10^{-3}
Maximum Degree	3, 992	5,706
Average Degree	65	146
Average Labels	1.4	1.3

Table 1. *Descriptive statistics of the Flickr data set, reflecting the degree distribution. Categories are groups, actors are users, and links are friendship ties (Tang & Liu, 2009).*

Analysis

To test our hypothesis that Flickr users follow the principle of homophily and have more friendship ties within their group than outside their group, we conducted a set of six experiments using Python and NetworkX (NetworkX, 2018). Python is a high-level, open-source programming language, and NetworkX is a Python package used to study graphs and networks. Due to the large size of our dataset, a powerful programming language and library capable of executing complex equations with high efficiency was required, making Python and NetworkX the most advantageous tools for testing our hypothesis. NetworkX uses scientific computing libraries written in C, which are optimized for such large datasets. Additionally, Python is a relatively user-friendly language with easily searchable code for any type of analysis (Hagberg, Shult, & Swart, 2008).

Before beginning our experiments, we needed to determine group membership. This information was found in the “group edges” file of the dataset, which lists all users and their corresponding group memberships. Flickr users each subscribe to an average of 1.3 groups (see Table 1). In order to properly test homophily, users must be assigned to only one group. If users remain in multiple groups, determining homophily becomes exceptionally challenging. Because our data is binary (i.e., users are either in group x or not in group x), weighted group scores regarding how active each user is in each group, or which group they prefer/identify with most strongly, are not provided. We considered two options: assigning users to the largest group for which they are subscribed or assigning them to the smallest group for which they are subscribed. It is assumed that the largest group is broad and related to more general interests (e.g., Puppies, Nature), and the smallest group is more specialized (e.g., Human Factors Graduate Students at California State University, Long Beach). Based on these assumptions combined with the average number of groups per user being 1.3, we were able to infer likely engagement within these groups. That is, users may be most inclined to participate in groups of general interest and/or groups of specific, specialized interest.

Once users were assigned to one group (largest or smallest), we determined how many members belong to each group and then ordered the groups in NetworkX from largest to smallest, followed by smallest to largest. We then conducted experiments to test for homophily within the largest groups, within the smallest groups, between the largest groups, between the smallest groups, and between the largest and smallest groups. In order to analyze and visualize the dataset as a whole, a degree assortativity analysis was conducted at the end.

Results

In order to obtain a comprehensive view of potential homophily within the Flickr dataset, six separate analyses were conducted in Python's NetworkX (Appendices A, B). Coefficients produced by homophily analyses can be negative numbers, positive numbers, or 0. A negative coefficient indicates a disassortative network where users have more friendships outside their group than inside their group. A positive coefficient indicates an assortative network where users have more friendships inside their group than outside their group. A coefficient of 0 indicates that there is no relationship between friendship ties and group membership and that the friendship ties appear more random. The subsequent outputs from each experiment were evaluated against the Pearson correlation coefficient scale (Hinkle, Wiersma, & Jurs, 1990). Using this scale, positive and negative values are determined to be very high, high, moderate, low, or negligible. No negative coefficients were produced.

Experiment/Groups	Coefficient	Interpretation
1. Largest	0.1269	Very Low
2. Smallest	0.0697	Negligible/Low
3. Two Largest	0.2123	Low
4. Two Smallest	0.9687	Very High
5. Largest and Smallest	0.1625	Low
6. Degree	0.0715	Negligible/Low

Table 2. *Summary of results from homophily experiments*

Experiment 1

For our first experiment, we assigned each user to the largest group they are a part of and tested for homophily within those larger assigned groups. A coefficient of 0.1269 was obtained. This indicates very low homophily, only slightly more than what would be expected by chance.

Experiment 2

For our second experiment, we assigned each user to the smallest group they are a part of and tested for homophily within the smaller groups. A coefficient of 0.0697 was obtained. This indicates a very low to negligible level of homophily, only slightly more than would be found by chance.

Experiment 3

Similar to Experiment 1, we assigned each user to the largest group they are a part of. We then compared the two largest groups to each other and obtained a coefficient of 0.2123. This indicates low homophily (Figure 2).

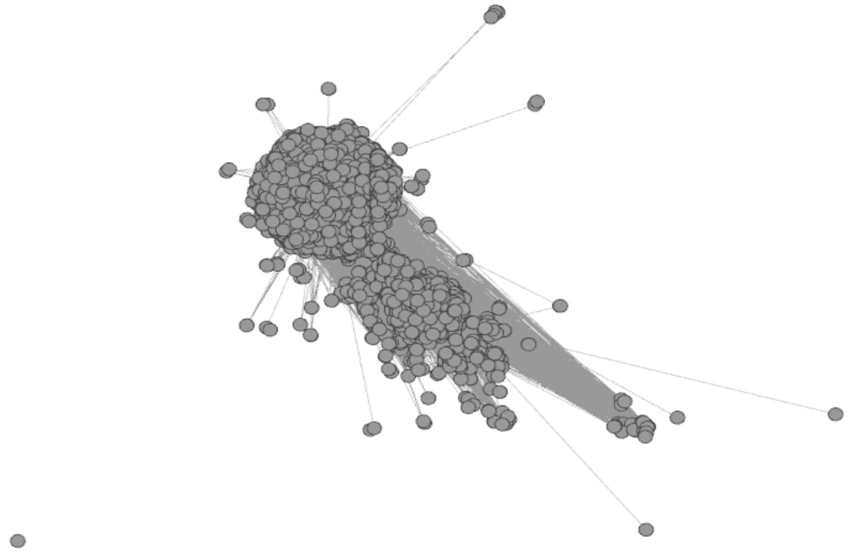


Figure 2. *Visualization of friendship ties between the two largest groups.*

Experiment 4

Similar to Experiment 2, we assigned each user to the smallest group they are a part of. We then compared the two smallest groups to each other and obtained a coefficient of 0.9687. This indicates very high homophily (Figure 3).

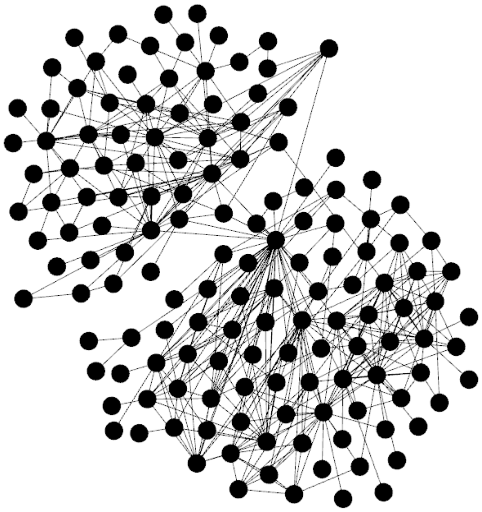


Figure 3. *Visualization of friendship ties between the two smallest groups.*

Experiment 5

Similar to Experiment 3, we assigned each user to the largest group they are a part of. We then compared the largest group and the smallest group and obtained a homophily coefficient of 0.1625. This indicates low homophily (Figure 4).

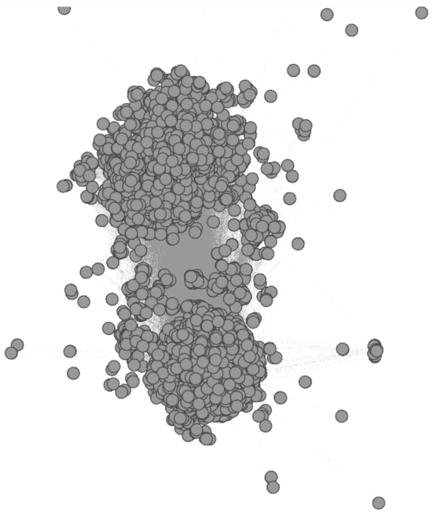


Figure 4. *Visualization of the friendship ties between the largest and smallest groups.*

Experiment 6

For our final experiment, we wanted to examine the graph as a whole. We ran a degree analysis on the entire dataset to determine if homophily occurs between high-degree nodes and other high-degree nodes, as well as between low-degree nodes and other low-degree nodes. A coefficient of 0.0715 was obtained. This indicates very low homophily, just slightly more than what would be found by chance. A low homophily coefficient also indicates that the high-degree nodes are not clustered in the center of the graph, but rather spread out (Figure 5).

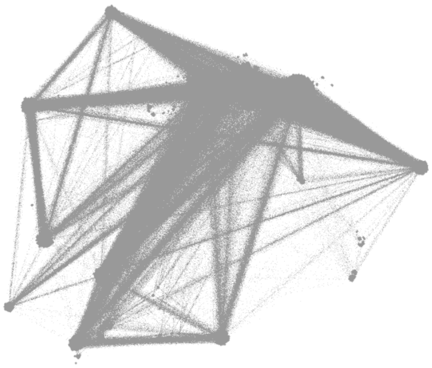


Figure 5. *Visualization of the friendship ties across the entire network. High-degree nodes are spread throughout the network, rather than clustered in the center.*

Because all coefficient values obtained are positive numbers, we can conclude that our hypothesis is supported, though homophily values are very low to negligible. Because the coefficient values are small, Flickr users tend to have ties more within their groups only slightly more than would be found by chance.

Conclusion

Flickr provides individuals with a sense of belonging and unity among others who share common interests in photography and videography. Because Flickr consists of groups of users who search or share content of similar nature, it was hypothesized that there would be homophily among the groups. The current study investigated homophily in six different experimental groups. It was determined that homophily accounted for the associations users created in the groups only slightly more than associations that would form by chance. This may be due to the online community nature of Flickr, wherein users can be a part of multiple groups more easily than they would in person due to proximity between individuals not being an issue.

Flickr users are joined together in groups based on shared interests. Because our dataset is lacking thorough, non-binary attributes to indicate levels of engagement in groups, or how friendship ties are formed, we are left to speculate as to why a user is in the group they are in, and how they choose friendships. For instance, both Experiments 2 and 4 analyzed the smallest groups in the network, wherein Experiment 2 tested for homophily within the smallest groups users subscribed to, and Experiment 4 tested for homophily between the two smallest groups in the network. Experiment 2 had the lowest homophily score (0.0697) and Experiment 4 had the highest (0.9687). One explanation for these opposing findings is that a high homophily score does not always indicate high density or saturation. That is, users do not necessarily prefer friendships inside their groups over friendships outside their groups, but rather that the two smallest groups, specifically, do not have significant overlap. Because it is assumed that the smaller groups are more specific and specialized, it is possible that their specializations are polar opposites (e.g., republican group vs. democrat group) where users typically choose one group over the other, rather than subscribing to both.

Another interesting finding concerning vastly contrasting scores is the comparison of Experiment 3 with Experiment 4. Experiment 3 tested for homophily between the two largest groups, and Experiment 4 (as mentioned above) tested for homophily between the two smallest groups. Experiment 3 had a much lower homophily score (0.2123) than Experiment 4 (0.9687). Because the two largest groups contain a significantly higher portion of Flickr users than the two smallest groups, and the content may be more general and broad, there is much more overlap between the two largest groups.

When examining the network as a whole, we found that Experiment 6 had a very low degree score (0.0715), meaning that high-degree users are likely to be friends with other high-degree users, and low-degree users are more likely to be friends with other low-degree users, only slightly more than would be expected by chance. High-and low-degree users can be mutual friends, though there may be a slight bias toward high-degree users choosing other high degree users, and low-degree users choosing other low-degree users. Interestingly, the low degree score and corresponding sociogram indicate that the high-degree users are more spread out among the network, rather than being densely clustered in the center of the network with the smaller nodes on the periphery.

Overall, the larger groups tend to have slightly higher homophily scores. This may be due to the diversity among users in the larger groups, and the sheer number of users subscribed to the largest groups (approximately 17% of all the nodes in the network are in the largest group). It stands to reason that a significant number of users are friends with members in those groups. Additionally, the larger groups may have such a high number of members because the subject matter is more general and appeals to greater, more diverse groups of users, while the smaller groups may be more specific in that they appeal to a small number of users. Hence, the slightly higher homophily scores for the largest groups may be due to the general content appealing to a greater, more diverse group of users.

From an organizational standpoint, Flickr could encourage users to build more friendships inside of their groups by increasing user engagement and supporting smaller, more specialized groups. To take a more democratic approach and encourage the fostering of friendships outside of the users' groups, Flickr could also advertise or promote groups that may be of interest to the user based on groups they are currently a part of.

Limitations

One major limitation of this analysis is that this Flickr dataset does not include the weighted scores used to illustrate user engagement and preference for particular groups. Users may be members of multiple groups, but not actively and equally involved in them all. This is problematic because including and analyzing user preference may have resulted in higher homophily scores. Additionally, having the ability to analyze the strength of friendship ties among users within groups for which they are most active could have boosted homophily scores as well. In addition to weighted group scores, the dataset would need to include descriptions of each group, rather than an anonymized list. By not including group descriptions, it is difficult to determine why users may choose membership in one group over others. Furthermore, despite our large dataset, it was only a subset of a much larger network that we did not have access to. Analyzing the full dataset may have provided us with more significant results. Future studies should address these concerns by utilizing the full dataset (or at least an additional randomly generated sample) that includes weighted group scores and group descriptions, as well as a measure for examining the strength of the ties within groups, and determining how such ties may be affected by the content of the group.

References

Dotan, Amir & Zaphiris, Panayiotis. (2010). A cross-cultural analysis of Flickr users from Peru, Israel, Iran, Taiwan and the UK. International Journal of Web Based Communities. 6. 284-302. 10.1504/IJWBC.2010.033753.

Faralli, S., Stilo, G., & Velardi, P. (2015). Large Scale Homophily Analysis in Twitter Using a Twixonomy. Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015).

Flickr. (n.d.). About Flickr. Retrieved May 1, 2018, from https://www.flickr.com/about

Hagberg, A., Shult, D., & Swart, P. (2008, August). Exploring network structure, dynamics, and function using NetworkX. *Proceedings of the 7th Python in Science Conference.* Pasadena, CA, 22-15.

Hinkle, D., Wiersma, W., & Jurs, S. (1990). Book Reviews: Applied Statistics for the Behavioral

Sciences. Boston: Houghton Mifflin Co., 1988. xix 682 pp. *Journal of Educational Statistics,15*(1), 84-87. doi:10.3102/10769986015001084

NetworkX. (2018, January). Retrieved from https://networkx.github.io/

Tang, L., & Liu, H. (2009). Relational learning via latent social dimensions. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 09. doi:10.1145/1557019.1557109

Tang, L., & Liu, H. (2009). Social Computing Data Repository at ASU [http://socialcomputing.asu.edu]. Tempe, AZ: Arizona State University, School of Computing, Informatics and Decision Systems Engineering.

Appendix A
Assortativity in Python

```
import os
import json
import networkx as nx
import pygraphviz as pgv
from networkx.drawing.nx_pydot import write_dot
from itertools import groupby, dropwhile

with open('groups.csv') as groups_input:
	groups = groups_input.read().splitlines()

with open('nodes.csv') as nodes_input:
	nodes = nodes_input.read().splitlines()

with open('edges.csv') as edges_input:
	edges = [tuple(edge.split(",")) for edge in edges_input.read().splitlines()]

with open('group-edges.csv') as group_edges_input:
	group_edges = [tuple(group_edge.split(",")) for group_edge in
group_edges_input.read().splitlines()]

# find each user's groups
group_edges = sorted(group_edges, key=lambda group_edge: group_edge[0])
user_to_groups = {}
```

```
for user, user_group_edges in groupby(group_edges, lambda group_edge: group_edge[0]):
        user_to_groups[user] = set([user_group_edge[1] for user_group_edge in
user_group_edges])

#print(len(user_to_groups))

# find distribution of user group membership count
user_group_count = {}
for user, group in user_to_groups.items():
        group_count = len(user_to_groups[user])
        user_group_count[group_count] = user_group_count.get(group_count, 0) + 1

#print(user_group_count)

# find the count of users for each group
group_edges = [tuple(reversed(group_edge)) for group_edge in group_edges]
group_edges = sorted(group_edges, key=lambda group_edge: group_edge[0])

group_user_count = {}
for group, users in groupby(group_edges, lambda group_edge: group_edge[0]):
        group_user_count[group] = len(list(users))

# make a list of groups in descending user count order
group_user_counts = group_user_count.items()
group_user_counts = sorted(group_user_counts, key=lambda guc: guc[1], reverse=True)


# *** experiments ***
# uncomment one of the following to modify the graph for the appropriate experiment

# assign users to their smallest group
# group_user_counts = list(reversed(group_user_counts))

# only use nodes that have a single group assignment
# nodes = [node for node in nodes if len(user_to_groups[node]) == 1]

# only use nodes in the largest two groups
#group_user_counts = group_user_counts[:2]
#nodes = [node for node in nodes if group_user_counts[0][0] in user_to_groups[node] or
group_user_counts[1][0] in user_to_groups[node]]

# only use nodes in the largest and smallest groups
#group_user_counts = [group_user_counts[0], group_user_counts[-1]]
#nodes = [node for node in nodes if group_user_counts[0][0] in user_to_groups[node] or
group_user_counts[1][0] in user_to_groups[node]]
```

```
# only use nodes in the two smallest groups
# group_user_counts = [group_user_counts[-2], group_user_counts[-1]]
# nodes = [node for node in nodes if group_user_counts[0][0] in user_to_groups[node] or
group_user_counts[1][0] in user_to_groups[node]]

G = nx.DiGraph()
for node in nodes:
   # assign user to one and only one group
        if node in user_to_groups:
                group = next((group_user_count[0] for group_user_count in group_user_counts if
group_user_count[0] in user_to_groups[node]), None)
        else:
                group = None
        G.add_node(node, group=group)

print("graph has {} nodes".format(len(G)))

undirected_edges = list()
for (a, b) in edges:
        # add undirected input as directed edges as required by networkx
        if a in G and b in G:
           G.add_edge(a, b)
           G.add_edge(b, a)
           undirected_edges.append((a,b))

print("graph has {} directed edges".format(G.number_of_edges()))

print("writing graph dot file")
with open('graph.dot', 'w') as f:
        f.write('graph  {\n')
        for node in G.nodes():
                f.write('{};\n'.format(node))
        for (a, b) in undirected_edges:
                f.write('{} -> {};\n'.format(a, b))
        f.write('}\n')

print("running attribute_assortativity_coefficient")
r = nx.attribute_assortativity_coefficient(G, 'group')
print("coefficient is {}".format(r))
```

Appendix B
Degree Assortativity in Python

```
import os
import networkx as nx

with open('groups.csv') as groups_input:
	groups = groups_input.read().splitlines()

with open('nodes.csv') as nodes_input:
	nodes = nodes_input.read().splitlines()

with open('edges.csv') as edges_input:
	edges = [tuple(edge.split(",")) for edge in edges_input.read().splitlines()]

with open('group-edges.csv') as group_edges_input:
	group_edges = [tuple(group_edge.split(",")) for group_edge in
group_edges_input.read().splitlines()]

print(len(groups))
print(len(nodes))
print(len(edges))
print(len(group_edges))

G = nx.DiGraph()
G.add_nodes_from(nodes)
G.add_edges_from(edges)
```

```
for (a, b) in edges:
        G.add_edge(b, a)

r = nx.degree_assortativity_coefficient(G)
print(r)
```