Appendix. Electronic Companion for Building a Location-Based Set of Social Media Users

In this E-Companion we provide additional data analysis for the manuscript "Building a Location-Based Set of Social Media Users".

A. Proof of Theorem 1

Before proving Theorem 1, it is useful to establish our notation and provide an important Lemma. We let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be the Energy Graph representation of for an energy function of the form in equation (3), which adheres to the assumptions in Section 3.2, where $\mathcal{V}$ is the set of nodes and $\mathcal{E}$ is the set of edges. A valid s-t cut is a partition of $\mathcal{V}$ into two subsets: S and T, where source node $s \in S$ and sink node $t \in T$. We define the cut-set $C \subset \mathcal{E}$ as the set of directed edges going from any node in set S to any node in set T. For arbitrary location classifications $\mathbf{L} = (\ell_1, \ell_2, \ldots, \ell_N) \in \{0,1\}^N$, consider the s-t cut on the Energy Graph $\mathcal{G}$ that partitions the nodes according to their location classes. We denote the set of nodes in the subset belonging to the source node as $S_{\mathbf{L}}$, where node $u_i \in S_{\mathbf{L}}$ for all users i for which $\ell_i = 1$. Likewise, $u_j \in T_{\mathbf{L}}$ for all j such that $\ell_j = 0$. We refer to this cut as the $\mathbf{L}$-*configuration* cut on the Energy Graph.

We state our Lemma.

Lemma 1 **(Graph Equivalence)**. *Suppose we are given an energy function of the form*

$$E(\mathbf{L}) = \sum_i \phi(\mathbf{x}_i, \ell_i) + \sum_{i<j} \psi(\mathbf{z}_{i,j}, \ell_i, \ell_j),$$

where functions $\phi(\mathbf{x}_i, \ell_i)$ and $\psi(\mathbf{z}_{i,j}, \ell_i, \ell_j)$ adhere to the assumptions in Section 3.2. Then, for arbitrary location classification vector $\mathbf{L} \in \{0,1\}^N$, the value of the function $E(\mathbf{L})$ is equal to the capacity of the $\mathbf{L}$-configuration cut on the corresponding energy graph.

Proof of Lemma 1. Let $G = (\mathcal{V}, \mathcal{E})$ be the energy graph representation of energy function E, and consider an arbitrary fixed location classification vector $\mathbf{L} \in \{0,1\}^N$. From our definition of the $\mathbf{L}$-configuration cut, $\ell_i = 1$ implies node $u_i \in S_{\mathbf{L}}$ and $(u_i, t) \in C_{\mathbf{L}}$. Likewise, $\ell_i = 0$ implies $(S, u_i) \in C_{\mathbf{L}}$. Now consider an arbitrary pair of user nodes $u_i, u_j,\ i \neq j$. One of the following cases apply: (Case 1) both users are assigned to location class 1 ($\ell_i = \ell_j = 1$), (Case 2) one of the users is in location class 1 and the other is assigned to location class 0 ($\ell_i \neq \ell_j$), or (Case 3) both users are assigned to location class 0 ($\ell_i = \ell_j = 0$). We consider the implications of each case on cut-set $C_{\mathbf{L}}$.

Case 1. Because u_i and u_j are both in set $S_{\mathbf{L}}$, edges between these two nodes are not in $C_{\mathbf{L}}$. It follows that

$$\ell_i = \ell_j = 1 \Rightarrow (u_i, u_j), (u_j, u_i) \notin C_{\mathbf{L}}.$$

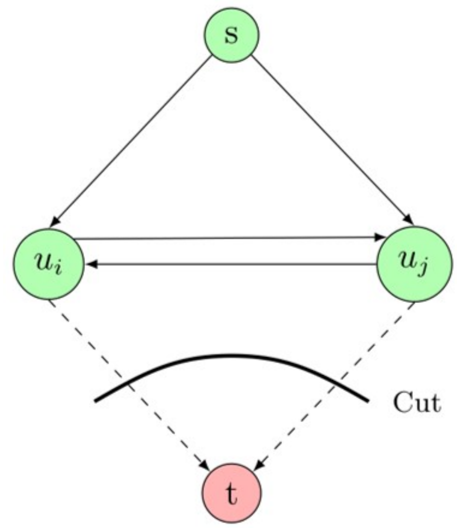


Figure 11 Illustration of Case 1 cut. Nodes in set $S_{\mathbf{L}}$ are shaded green, while nodes in set $T_{\mathbf{L}}$ are shaded red. Dashed edges are in cut-set $C_{\mathbf{L}}$.

However, edges (u_i, t) and (u_j, t) are in the cut-set $C_{\mathbf{L}}$. Figure 11 provides an illustration of this case.

Case 2. Without loss of generality, assume $\ell_i = 1$ and $\ell_j = 0$, implying $(u_i, t) \in C_{\mathbf{L}}$ and $(s, u_j) \in C_{\mathbf{L}}$. Because $u_i \in S$ and $u_j \in T$, edge (u_i, u_j) is also in the cut-set $C_{\mathbf{L}}$. This case is depicted in Figure 12.

Note that the reverse edge from u_j to u_i in Figure 12 is *not* in the cut-set because they go from set $T_{\mathbf{L}}$ to set $S_{\mathbf{L}}$.

Case 3. Finally, we consider the case in which $\ell_i = \ell_j = 0$. This case is very similar to Case 1: edges (s, u_i) and (s, u_j) are in $C_{\mathbf{L}}$, while other edges incident to these nodes are not in the cut-set (see Figure 13).

From these rules we can identify all of the edges comprising set cut-set $C_{\mathbf{L}}$ for an arbitrary location class vector $\mathbf{L}$. The total capacity of the cut is the sum of all of these edge capacities:

$$\begin{aligned}
\sum_{a \in C_{\mathbf{L}}} c_a &= \underbrace{\sum_{i:\, \ell_i = 0} c_{(s,u_i)} + \sum_{i:\, \ell_i = 1} c_{(u_i,t)}}_{\mathbf{L}-\text{configuration}} + \underbrace{\sum_{i<j:\, \ell_i=1,\, \ell_j=0} c_{(u_j,u_i)} + \sum_{i<j:\, \ell_i=0,\, \ell_j=1} c_{(u_i,u_j)}}_{\text{Case 2 user node links}} \\
&= \sum_{i:\, \ell_i=0} \left(\phi(\mathbf{x}_i, 0) + \frac{1}{2} \sum_{j:\, \neq i} \psi(\mathbf{z}_{i,j}, 0, 0) \right) + \sum_{i:\, \ell_i=1} \phi(\mathbf{x}_i, 1) \\
&\quad + \sum_{i<j:\, \ell_i=1,\, \ell_j=0} \left(\psi(\mathbf{z}_{i,j}, 1, 0) - \frac{1}{2} \psi(\mathbf{z}_{i,j}, 0, 0) \right) + \sum_{i<j:\, \ell_i=0,\, \ell_j=1} \left(\psi(\mathbf{z}_{i,j}, 0, 1) - \frac{1}{2} \psi(\mathbf{z}_{i,j}, 0, 0) \right) \\
&= \sum_{i=1}^{N} \phi(\mathbf{x}_i, \ell_i) + \sum_{i<j:\, \ell_i=\ell_j=0} \psi(\mathbf{z}_{i,j}, 0, 0) + \sum_{i<j:\, \ell_i=1,\, \ell_j=0} \psi(\mathbf{z}_{i,j}, 1, 0) + \sum_{i<j:\, \ell_i=0,\, \ell_j=1} \psi(\mathbf{z}_{i,j}, 0, 1)
\end{aligned}$$

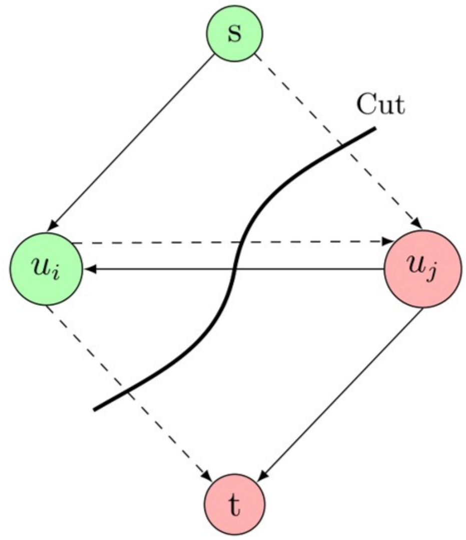


Figure 12 Illustrations of both minimum cut possibilities for Case 2. Nodes in set $S_{\mathbf{L}}$ are shaded green, while nodes in set $T_{\mathbf{L}}$ are shaded red. Dashed edges are in cut-set $C_{\mathbf{L}}$.

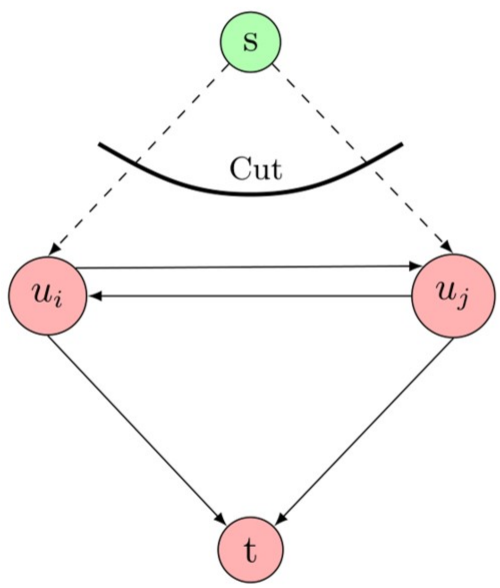


Figure 13 Illustration of Case 3 cut. Nodes in set $S_{\mathbf{L}}$ are shaded green, while nodes in set $T_{\mathbf{L}}$ are shaded red. Dashed edges are in cut-set $C_{\mathbf{L}}$.

$$= \sum_{i=1}^{N} \phi(\mathbf{x}_i, \ell_i) + \sum_{i<j} \psi(\mathbf{z}_{i,j}, \ell_i, \ell_j) = E(\mathbf{L}). \quad \square$$

Proof of Theorem 1. Theorem 1 follows almost immediately from Lemma 1. Suppose we find the minimum s-t cut on the Energy Graph corresponding to energy function $E(\mathbf{L})$, and let sets S

and T be the corresponding partition of $\mathcal{V}$ and $C \subset \mathcal{E}$ be the cut-set. In order for the cut to be valid, each user node u_i must be in either set S or set T. For each user node $u_i \in S$, edge (u_i, t) is in the cut-set C and we set $\ell_i = 1$. For each user node $u_i \in T$, edge $(s, u_i) \in C$ and we set $\ell_i = 0$. Let $\mathbf{L}^\star$ be the resulting vector of location assignments.

From Lemma 1, the capacity of this cut is $E(\mathbf{L}^\star)$. Because the sets S and T were constructed from the minimum capacity s-t cut on the graph, there cannot be another location assignment $\mathbf{L}$ for which $E(\mathbf{L}) < E(\mathbf{L}^\star)$, as such a vector would allow for the construction of an s-t cut with lower capacity. $\square$

B. Sensitivity Analysis

B.1. Corinto, Colombia

In this section we discuss and illustrate the sensitivity of the performance of the expand-classify approach on Corinto, Columbia to the inputs γ, λ, as well as the logistic regression regularization. The parameter γ serves as the magnitude of the sigmoid curve that governs the decay of link energy as the number of friends and followers increases. Higher values of γ result in larger link energies, which cause network connections to have more influence over each user's classification.

Figure 14 depicts the classification ROC curve plotted using several different values for γ. Of particular note is the case where $\gamma = 0$, which recovers the logistic regression classification without any network information. Comparison with this curve provides a quantification of the utility of the network structure in this classification model. Based on the AUC metric, we find that optimal performance appears to occur for higher values of γ, with $\gamma = \log(10)$ producing an AUC of 0.94. However, small variations from this value do not appear to substantially impact performance. Using only the logistic regression model ($\gamma = 0$) produces an AUC of approximately 0.64, showing that accounting for network connections in the model substantially improves classification performance.

The parameter $\lambda \in [0, 1]$ sets the link energy for users that are both classified in location set 0 (outside of the target location). We have set this parameter to 0.98, so that these relationships are approximately the same cost as relationships for which one user is in the target location and the other is not. Figure 14 depicts the sensitivity of the ROC curve to this value. Higher values of λ appear to produce the best results, and very low values performing very poorly. This poor performance results from users in the target location being misclassified as location 0 as a result of relationships with other users in location class 0. Good performance is maintained for values of $\lambda > 0.75$.

Finally, we investigate the model sensitivity to the logistic regression regularization. In a similar fashion to the above analyses, we fit $L1$ and $L2$ regularized logistic regression models to the geo-located training data using different value for the regularization coefficient. Without fitting the

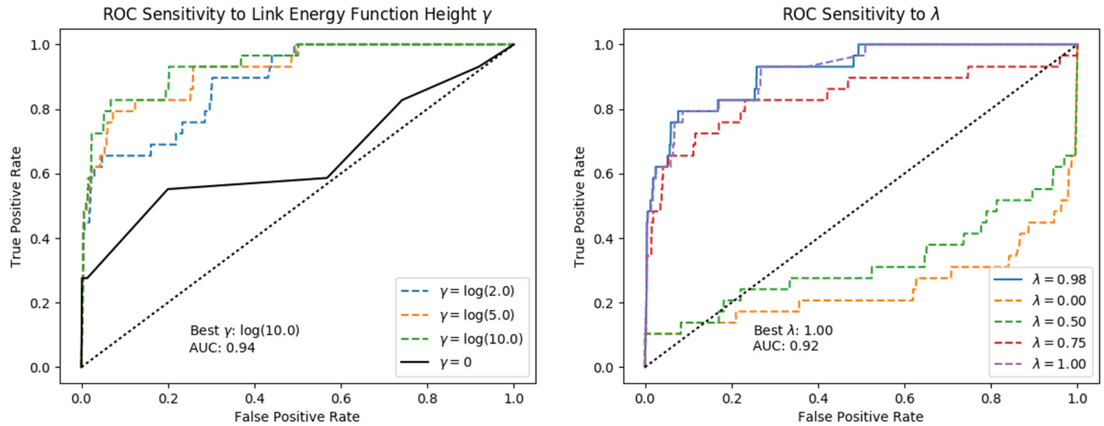
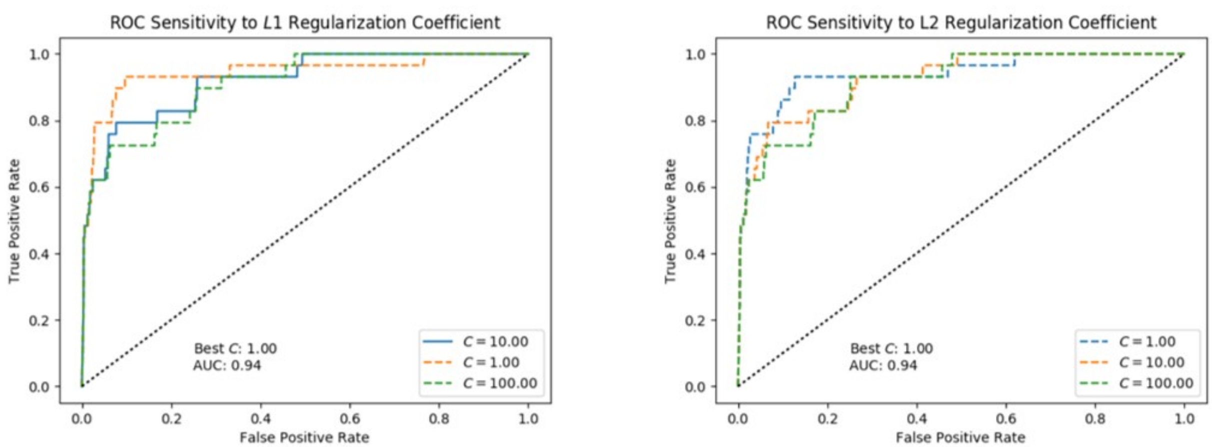


Figure 14 (left) Sensitivity of Corinto user classification to parameter γ. (right) Sensitivity of Corinto user classification to parameter λ.

regularization coefficient through model validation, we applied the resulting linear function directly as the profile energy model. We found that the model performance was neither sensitive to the regularization norm ($L1$ vs. $L2$) nor to the regularization coefficient, except for in cases in which we significantly over-regularized the logistic regression. Figure 15 shows the classification ROC using for several regularization constants for both $L1$ and $L2$ regularization. We observe that a lower value of C, which implies a more regularized model, results in a slight increase in performance from the value found through validation.



$L1$ Regularization $L2$ Regularization

Figure 15 Sensitivity of Corinto user classification to logistic regression regularization.

B.2. Casimiro de Abreu, Brazil

Figure 16 shows the sensitivity of the ROC for Casimiro de Abreu, Brazil on γ and λ. The results do not appear to be very sensitive to the value of γ, but the AUC decreases for lower values of λ. Larger values of λ also seem to show slightly better performance.

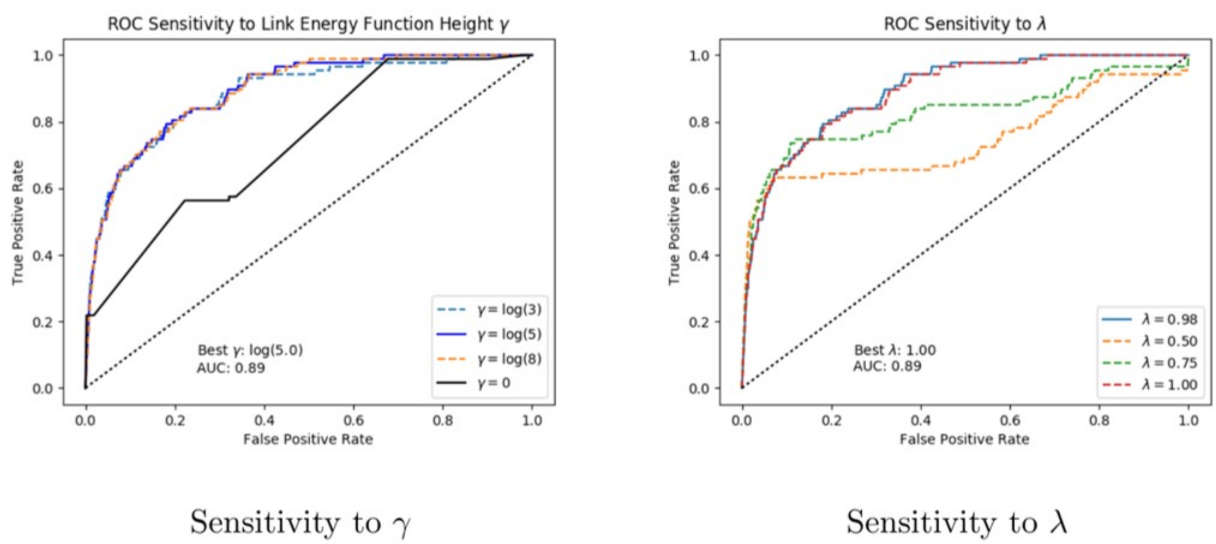


Figure 16 Results and Sensitivity of Casimiro de Abreu user classification.

Note that the performance of the fixed profile energy function as a classifier is also plotted in this Figure for $\gamma = 0$, indicated by a solid black line. The AUC of this classifier is 0.74.

B.3. Caracas, Venezuela

Figure 17 shows ROC curves for Caracas, Venezuela for different values of γ. The best performance is achieved for $\gamma = 0$, which corresponds classify the users using logistic regression only with no network information.

C. World Cities Data

The following cities were used in set W_2 in our specific location collection implementations. This list is extracted from the World Cities Dataset created and maintained by MaxMind, available at http://www.maxmind.com/ (MaxMind.com 2017)

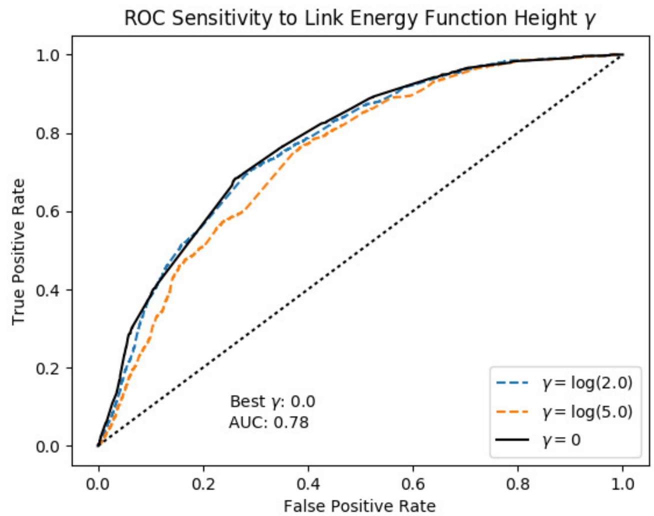


Figure 17 Caracas, Venezuela Performance.

dubai	kabul	yerevan	luanda	cordoba	rosario
vienna	adelaide	brisbane	melbourne	perth	sydney
baku	dhaka	khulna	brussels	ouagadougou	sofia
belem	belo horizonte	brasilia	campinas	curitiba	fortaleza
goiania	guarulhos	manaus	nova iguacu	porto alegre	recife
rio de janeiro	salvador	sao paulo	minsk	montreal	toronto
vancouver	kinshasa	lubumbashi	brazzaville	abidjan	santiago
douala	yaounde	anshan	changchun	chengdu	chongqing
dalian	datong	fushun	fuzhou	guangzhou	guiyang
handan	hangzhou	harbin	hefei	huainan	jilin
jinan	kunming	lanzhou	luoyang	nanchang	nanjing
peking	qingdao	rongcheng	shanghai	shenyang	shenzhen
suzhou	taiyuan	tangshan	tianjin	urumqi	wuhan
wuxi	xian	xianyang	xinyang	xuzhou	barranquilla
bogota	cali	medellin	prague	berlin	hamburg
munich	copenhagen	santo domingo	algiers	guayaquil	quito
alexandria	cairo	gizeh	barcelona	madrid	addis abeba
paris	london	tbilisi	accra	kumasi	conakry
port-au-prince	budapest	bandung	bekasi	depok	jakarta
makasar	medan	palembang	semarang	surabaya	tangerang
dublin	agra	ahmadabad	allahabad	amritsar	aurangabad
bangalore	bhopal	bombay	calcutta	delhi	faridabad
ghaziabad	haora	hyderabad	indore	jabalpur	jaipur
kalyan	kanpur	lakhnau	ludhiana	madras	nagpur
new delhi	patna	pimpri	pune	rajkot	surat
thana	vadodara	varanasi	visakhapatnam	baghdad	esfahan
karaj	mashhad	qom	shiraz	tabriz	milan
rome	hiroshima	kawasaki	kobe	nagoya	saitama
tokyo	nairobi	phnum penh	seoul	almaty	bayrut
beirut	tripoli	casablanca	fez	rabat	antananarivo
bamako	mandalay	rangoon	ecatepec	guadalajara	juarez
leon	mexico	monterrey	nezahualcoyotl	puebla	tijuana
kuala lumpur	maputo	benin	ibadan	kaduna	kano
lagos	maiduguri	port harcourt	managua	lima	davao
manila	faisalabad	gujranwala	hyderabad	karachi	lahore
multan	peshawar	rawalpindi	warsaw	bucharest	belgrade
chelyabinsk	kazan	moscow	nizhniy novgorod	novosibirsk	omsk
rostov-na-donu	saint petersburg	samara	ufa	volgograd	yekaterinburg
jiddah	mecca	riyadh	khartoum	umm durman	stockholm
singapore	freetown	dakar	mogadishu	aleppo	damascus
bangkok	adana	ankara	bursa	gaziantep	istanbul
izmir	kaohsiung	kaohsiung	taichung	taipei	dar es salaam
kiev	odesa	kampala	phoenix	los angeles	san diego
chicago	new york	philadelphia	dallas	houston	san antonio
montevideo	tashkent	caracas	maracaibo	valencia	hanoi
ha noi	ho chi minh city	cape town	durban	johannesburg	pretoria