Covariance Matrix Adaptation Evolution Strategy for Link Prediction in Dynamic Social Networks

Catherine A. Bliss, Morgan R. Frank, Christopher M. Danforth, & Peter Sheridan Dodds

Department of Mathematics & Statistics
Vermont Complex Systems Center
Computational Story Lab
Vermont Advanced Computing Core
University of Vermont

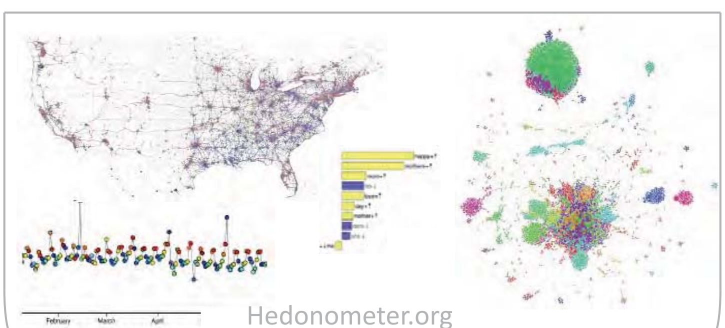
Background
Data
Reciprocal reply networks
Link prediction
Similarity indices
Evolutionary computation
Results
Conclusions

Background
Data
Reciprocal reply networks
Link prediction
Similarity indices
Evolutionary computation
Results
Conclusions

Background
Data
Reciprocal reply networks
Link prediction
Similarity indices
Evolutionary computation
Results
Conclusions

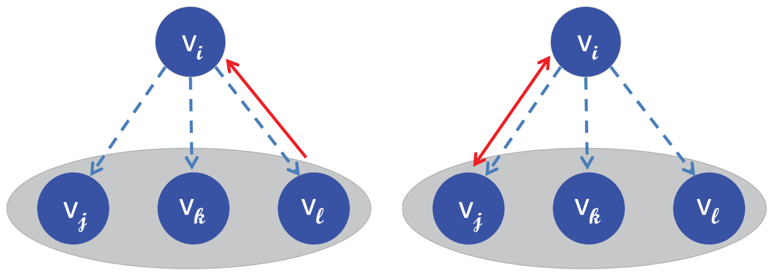
40,000 tweets (100MB) / min.
50 million tweets (150GB) / day
50 billion tweets (100TB) / 5+ years

Background
Data
Reciprocal reply networks
Link prediction
Similarity indices
Evolutionary computation
Results
Conclusions



40,000 tweets (100MB) / min.
50 million tweets (150GB) / day
50 billion tweets (100TB) / 5+ years

Reciprocal-reply networks



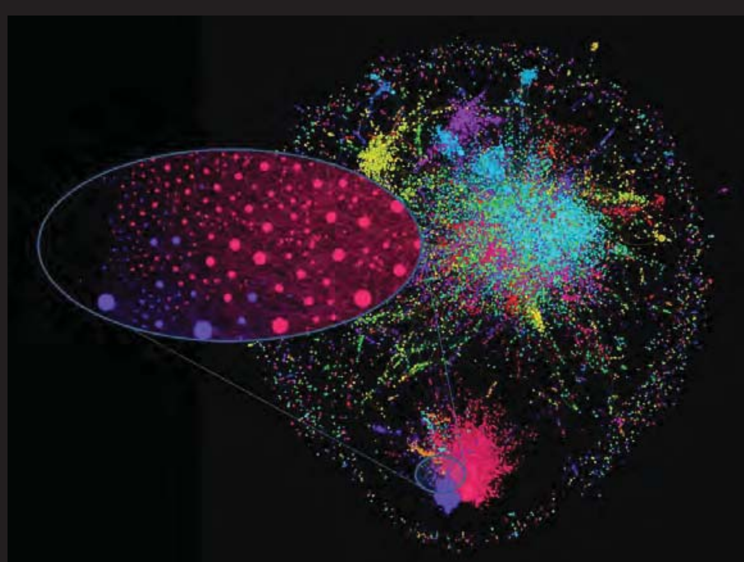
Unreciprocated reply   Reciprocated replies

We define a *reciprocal-reply network* as an unweighted, undirected network in which a link is established between nodes v_i and v_j if we observe reciprocal replies between these nodes **during the unit of time under analysis.**[1]

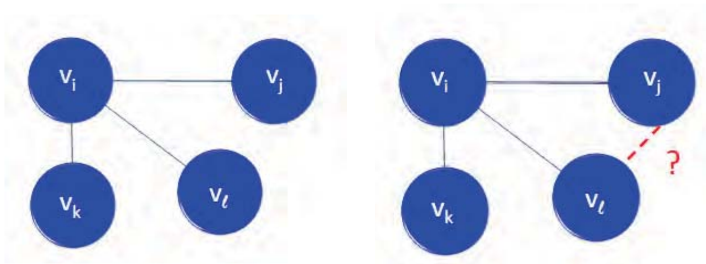
---

[1] C. A. Bliss, I. M. Kloumann, K. D. Harris, C. M. Danforth & P. S. Dodds. 2012. Twitter reciprocal reply networks exhibit assortativity with respect to happiness, Journal of Computational Science

Twitter RRN Link Prediction



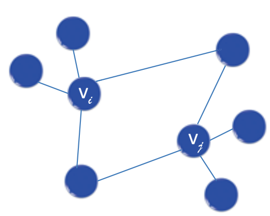
The link prediction problem: *Given a snapshot of the network at time=t, can we predict links which will appear in time=t+1?*



Time t **Time t+1**

Similarity indices

- Liben-Nowell & Kleinberg (2007) author collaboration networks ($N \propto 10^3$)
- Use similarity indices to rank the most likely occurring top N links

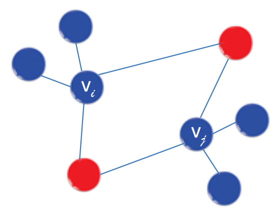


Common neighbors (Newman, 2001)

Background
Data
Reciprocal reply networks
Link prediction
Similarity indices
Evolutionary computation
Results
Conclusions

Similarity indices

- Liben-Nowell & Kleinberg (2007) author collaboration networks ($N \propto 10^3$)
- Use similarity indices to rank the most likely occurring top N links
  $C(u,v) = |\Gamma(u) \cap \Gamma(v)|$

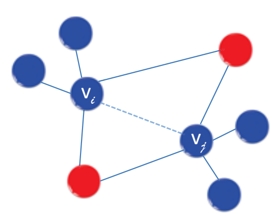


Common neighbors
(Newman, 2001)

Similarity indices

Background
Data
Reciprocal reply networks
Link prediction
Similarity indices
Evolutionary computation
Results
Conclusions

- Liben-Nowell & Kleinberg (2007) author collaboration networks ($N \propto 10^3$)
- Use similarity indices to rank the most likely occurring top N links
  $C(u,v) = |\Gamma(u) \cap \Gamma(v)|$



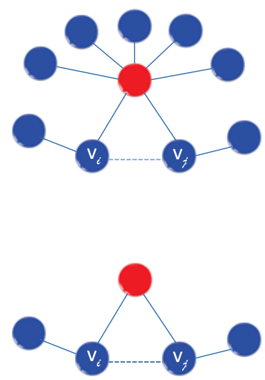
Common neighbors
(Newman, 2001)

Similarity indices

- Liben-Nowell & Kleinberg (2007) author collaboration networks ($N \propto 10^3$)
- Use similarity indices to rank the most likely occurring top N links

$$C(u,v) = |\Gamma(u) \cap \Gamma(v)|$$

$$R(u,v) = \sum_{z \in \Gamma(u) \cap \Gamma(v)} \frac{1}{|\Gamma(z)|}$$



Resource Allocation
(Zhou, Lu, & Zhang, 2009)

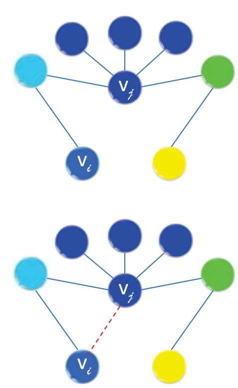
Similarity indices

- Liben-Nowell & Kleinberg (2007) author collaboration networks ($N \propto 10^3$)
- Use similarity indices to rank the most likely occurring top N links

$$C(u,v) = |\Gamma(u) \cap \Gamma(v)|$$

$$R(u,v) = \sum_{z \in \Gamma(u) \cap \Gamma(v)} \frac{1}{|\Gamma(z)|}$$

$$W(u,v) = 1 - \frac{1}{2} \sum |f_{u,n} - f_{v,n}|$$

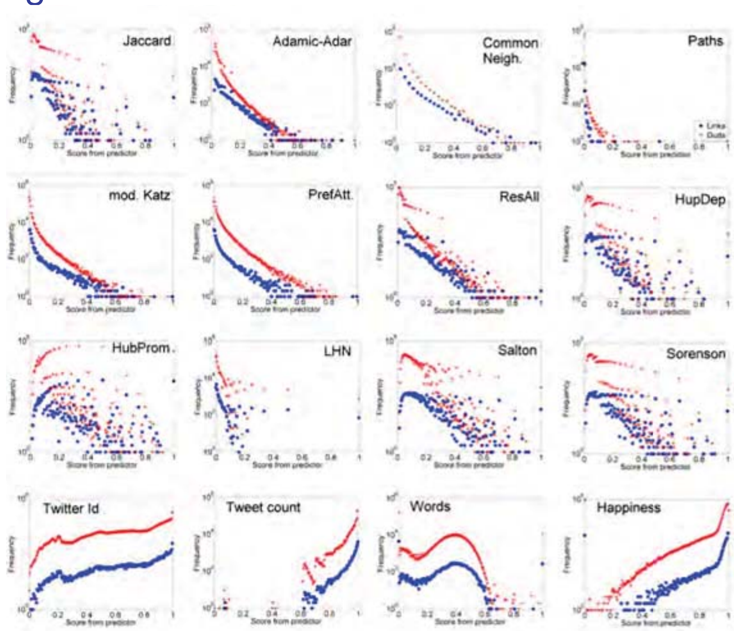


Word similarity
(Bliss et al., 2012)

Similarity indices

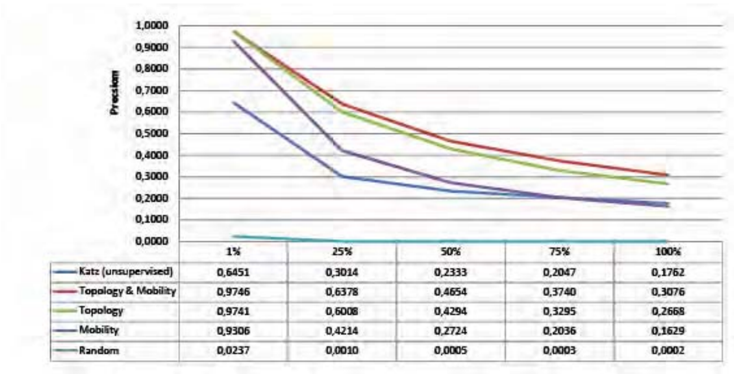
Common neighbors	$C(u,v) = \lvert\Gamma(u)\cap\Gamma(v)\rvert$
Jacard	$J(u,v) = \frac{\lvert\Gamma(u)\cap\Gamma(v)\rvert}{\lvert\Gamma(u)\cup\Gamma(v)\rvert}$
Adamic-Adar	$A(u,v) = \sum_{z\in\Gamma(u)\cap\Gamma(v)} \frac{1}{log(\lvert\Gamma(z)\rvert)}$
Pref Attachment	$Pr(u,v) = k_u \times k_v$
Hub promoted	$Hp(x,y) = \frac{\lvert\Gamma(u)\cap\Gamma(v)\rvert}{\min\{k_u,k_v\}}$
Hub depressed	$Hd(u,v) = \frac{\lvert\Gamma(u)\cap\Gamma(v)\rvert}{\max\{k_u,k_v\}}$
LHN	$L(u,v) = \frac{\lvert\Gamma(u)\cap\Gamma(v)\rvert}{k_u k_v}$
Salton	$Sa(u,v) = \frac{\lvert\Gamma(u)\cap\Gamma(v)\rvert}{\sqrt{k_u k_v}}$
Sorenson	$So(u,v) = \frac{2\lvert\Gamma(u)\cap\Gamma(v)\rvert}{k_u+k_v}$
Resource Allocation	$R(u,v) = \sum_{z\in\Gamma(u)\cap\Gamma(v)} \frac{1}{\lvert\Gamma(z)\rvert}$
Average Path Weight	$P(u,v) = \frac{\sum_{p\in\mathcal{P}_2(u,v)\cup\mathcal{P}_3(u,v)} w_p}{\lvert\mathcal{P}_2(u,v)\rvert+\lvert\mathcal{P}_3(u,v)\rvert}$
Katz	$K = \sum_{n=1}^{\infty} \beta^n A^n$
Tweet count similarity	$T(u,v) = 1 - \frac{\lvert T(u)-T(v)\rvert}{\max\{\lvert T(a)-T(b)\rvert\}_{a,b\in V}}$
Word similarity	$W(u,v) = 1 - \frac{1}{2}\sum_{n=1}^{50000} \lvert f_{u,n} - f_{v,n}\rvert$
Happiness similarity	$\mathrm{H}(u,v) = 1 - \frac{\lvert h(u)-h(v)\rvert}{\max\{\lvert h(a)-h(b)\rvert\}_{a,b\in V}}$
Id similarity	$I(u,v) = 1 - \frac{\lvert Id(u)-Id(v)\rvert}{\max\{\lvert Id(a)-Id(b)\rvert\}_{a,b\in V}}$

Signal detection



Background
Data
Reciprocal reply networks
Link prediction
Similarity indices
Evolutionary computation
Results
Conclusions

Combining indices



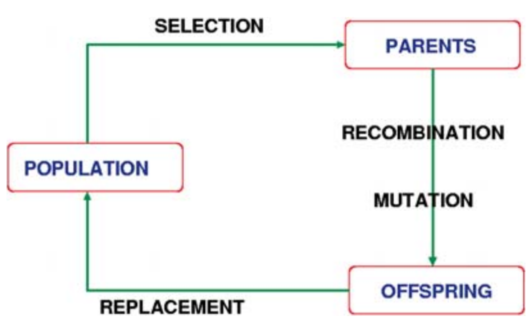
	1%	25%	50%	75%	100%
Katz (unsupervised)	0,6451	0,3014	0,2333	0,2047	0,1762
Topology & Mobility	0,9746	0,6378	0,4654	0,3740	0,3076
Topology	0,9741	0,6008	0,4294	0,3295	0,2668
Mobility	0,9306	0,4214	0,2724	0,2036	0,1629
Random	0,0237	0,0010	0,0005	0,0003	0,0002

Wang et al., (2011) examine mobile call graphs $N \propto 10^4$

- Combination of topological and node based similarity indices outperform single indices
- Use supervised learning
- Examine a small subset of node-node pairs, $\propto 10^3$
- Challenge - methods for large, sparse networks

Background
Data
Reciprocal reply networks
Link prediction
Similarity indices
Evolutionary computation
Results
Conclusions

Evolutionary computation



Background
Data
Reciprocal reply networks
Link prediction
Similarity indices
Evolutionary computation
Results
Conclusions

CMA-ES implementation[2]

Individual

An individual or candidate solution is a vector, $\vec{w} \in \boldsymbol{R}^{16}$.

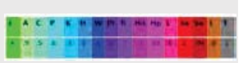
J	A	C	P	K	H	W	Pr	R	Hd	Hp	L	Sa	So	I	T
.4	.9	.5	.8	.3	.2	.1	.6	.7	-.1	.01	.6	.2	.04	.8	.1

[2] Hansen, N. (2006). The CMA Evolution Strategy: A Comparing Review. In J.A. Lozano, P. Larrañga, I. Inza and E. Bengoetxea (eds.). Towards a new evolutionary computation. Advances in estimation of distribution algorithms. pp. 75-102, Springer.

CMA-ES implementation[2]

Individual

An individual or candidate solution is a vector, $\vec{w} \in \boldsymbol{R}^{16}$.



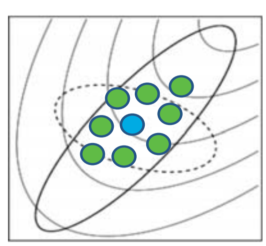
Reproduction & Mutation

CMA-ES

From 1 individual, generate a Gaussian cloud of candidate solutions in $\boldsymbol{R}^{16}$ using the covariance matrix.

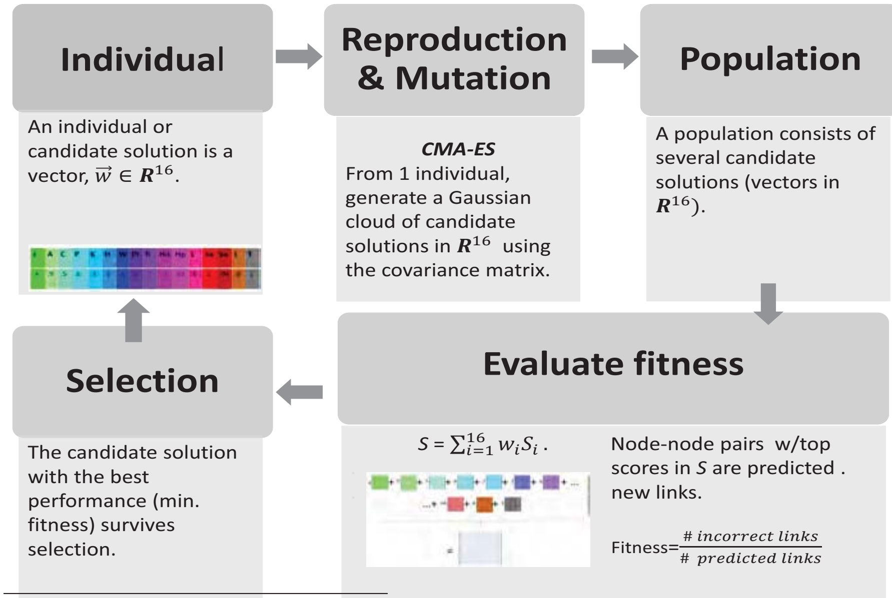
Population

A population consists of several candidate solutions (vectors in $\boldsymbol{R}^{16}$).



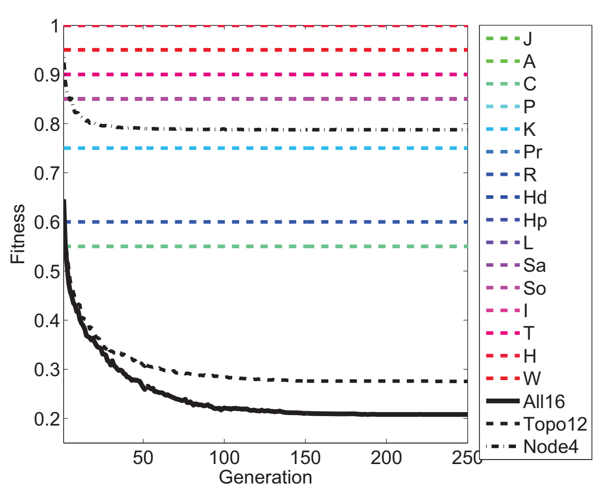
[2] Hansen, N. (2006). The CMA Evolution Strategy: A Comparing Review. In J.A. Lozano, P. Larrañga, I. Inza and E. Bengoetxea (eds.). Towards a new evolutionary computation. Advances in estimation of distribution algorithms. pp. 75-102, Springer.

CMA-ES implementation[2]

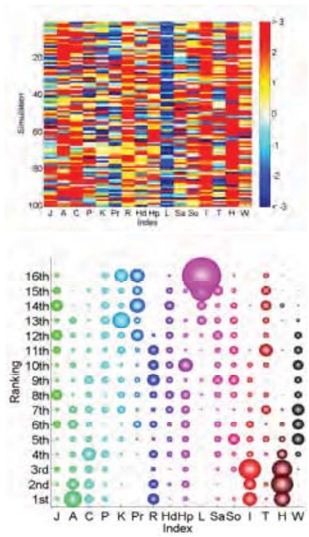


[2] Hansen, N. (2006). The CMA Evolution Strategy: A Comparing Review. In J.A. Lozano, P. Larrañga, I. Inza and E. Bengoetxea (eds.). Towards a new evolutionary computation. Advances in estimation of distribution algorithms. pp. 75-102, Springer.

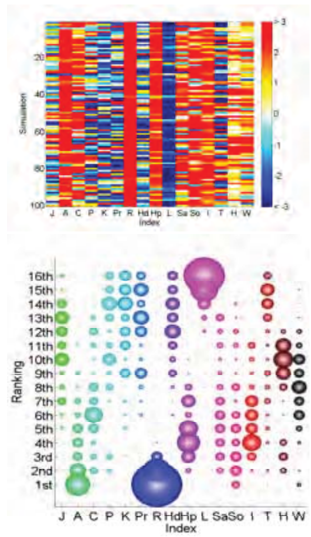
Fitness



Best predictors

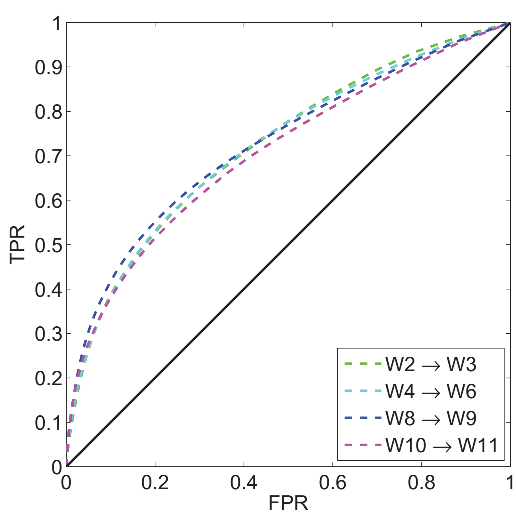


Week 7 to 8, top N=20



Week 7 to 8, top N=200

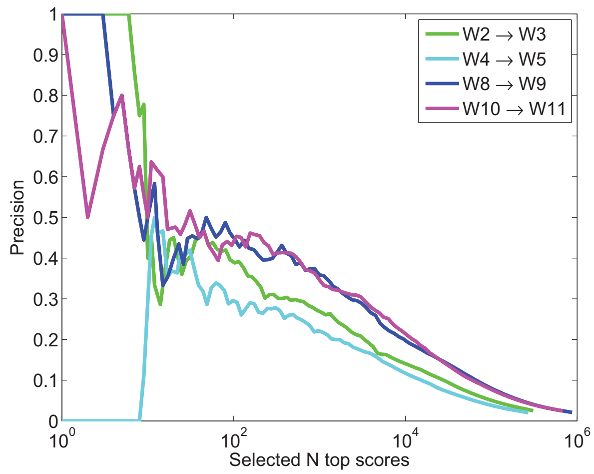
Receiver Operating Curve



ROC depicts $TPR = \frac{TP}{TP+FN}$ as a function of $FPR = \frac{FP}{FP+TN}$. The area under the curve (AUC) represents the chance that our predictor assigns a higher score to user-user pairs who exhibit new links than user-user pairs who do not exhibit new links. AUC>0.70

Background
Data
Reciprocal reply networks
Link prediction
Similarity indices
Evolutionary computation
Results
Conclusions

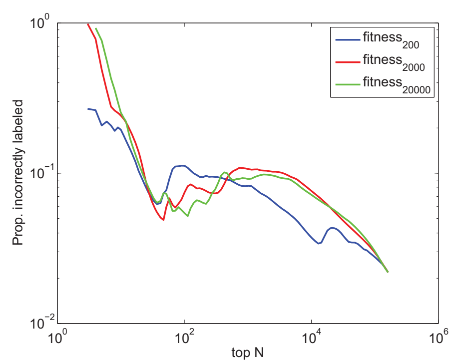
Precision



Precision depicts $\frac{TP}{TP+FP}$. High precision is achieved for top$N < 20$, which is often the region of interest. The precision for predicted links in $W4 \rightarrow W5'$ is lower than the other weeks and this may be due to missing data for those weeks

Missing data

- Sample 50% of tweets that we have
- Build nets & re-run CMA-ES on this smaller sample
- Compare # of links that are labeled false-positives to our fuller set of tweets
- Count # of FPs that are TPs (had we had seen the fuller set of data)



Background
Data
Reciprocal reply networks
Link prediction
Similarity indices
Evolutionary computation
Results
Conclusions

Conclusions

- Evolutionary algorithms show promise
- Many additional questions in link prediction (e.g., prediction of weights, prediction of link decay)
- Leveraging link prediction to understand network dynamics
- Further investigation of the role of incomplete data on network inference

Thank you

Background
Data
Reciprocal reply networks
Link prediction
Similarity indices
Evolutionary computation
Results
Conclusions

- **Manuscript**: In press at the *Journal of Computational Science*. Pre-print available at http://arxiv.org/abs/1304.6257
- **Contact**: www.cems.uvm.edu/~cabliss
- **Lab**: www.onehappybird.com