

*A Bayesian Approach Toward Finding  
Communities and Their Evolutions  
in Dynamic Social Networks*

Tianbao Yang<sup>1</sup> Yun Chi<sup>2</sup> Shenghuo Zhu<sup>2</sup> Yihong Gong<sup>2</sup> Rong Jin<sup>1</sup>

<sup>1</sup>Dept. CSE, Michigan State University, MI 48824, USA

<sup>2</sup>NEC Laboratories America, Cupertino, CA 95014, USA

May 01, 2009

# Outline

---

- Motivation and background
- Dynamic stochastic block model
- Bayesian inference and algorithm
- Experimental results
- Conclusion

# Outline

---

- Motivation and background
- Dynamic stochastic block model
- Bayesian inference and algorithm
- Experimental results
- Conclusion

# Motivation and background

---

- Analyzing communities in social networks is important
  - serves scientific purposes
    - e.g., sociology and social psychology
  - improves user experiences
    - e.g., friend recommendation services
  - provides business values
    - e.g., target advertisement, market segmentation analysis
- However, few studies focused on *dynamic communities*
  - macroscopic level: community structures may change
  - microscopic level: individuals' interests may change

# Our main contributions

- A novel dynamic stochastic block model (DSBM)
  - models communities and their evolutions in a unified probabilistic framework
- A Bayesian treatment for parameter estimation
  - instead of only the most likely values,
  - estimates the posterior distributions for unknown parameters
- A very efficient algorithm
  - executes in an incremental fashion to minimize the computational cost
  - fully takes advantage of the sparseness of data
  - both online learning and offline learning versions

# Outline

---

- Motivation and background
- Dynamic stochastic block model
- Bayesian inference and algorithm
- Experimental results
- Conclusion

# Dynamic stochastic block model — description

- An extension to the well-known Stochastic Block Model (SBM [HL76])
  - an *indicator variable*  $z$  for community membership
  - given  $z_i$  and  $z_j$ , link between  $i, j$  follows Bernoulli ( $P$ )
- An additional Markov chain ( $A$ ) for the transition probability among communities over consecutive time stamps

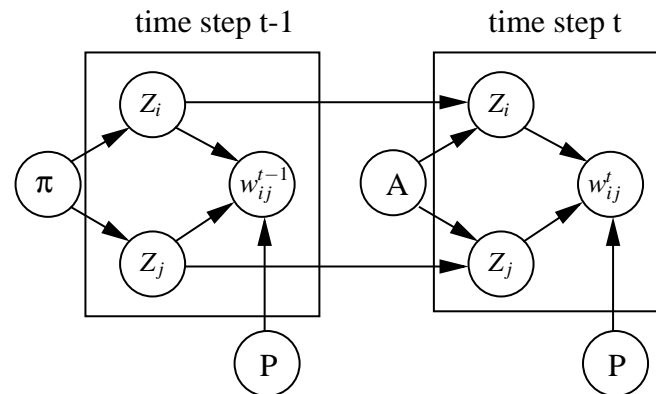


Figure 1: Graphical representation of DSBM

# Dynamic stochastic block model — description

---



Table 1: Generative Process of DSBM

---

For time 1:

generate the Social Network followed by SBM

For each time  $t > 1$ :

generate  $z_i^{(t)} \sim p(z_i^{(t)} | z_i^{(t-1)}, A)$

For each pair  $(i, j)$  at time  $t$ :

generate  $w_{ij}^{(t)} \sim \text{Bernoulli}(\cdot | P_{z_i^{(t)}, z_j^{(t)}})$

---



## Extensions

- can handle links of multinomial type
- can handle insertion and deletion of nodes

# Complete data likelihood

---

- Complete data likelihood

- $\Pr(\mathcal{W}_T, \mathcal{Z}_T | \pi, P, A) = \prod_{t=1}^T \Pr(W^{(t)} | Z^{(t)}, P) \prod_{t=2}^T \Pr(Z^{(t)} | Z^{(t-1)}, A) \Pr(Z^{(1)} | \pi)$

- Link emission probability

- $\Pr(W^{(t)} | Z^{(t)}, P) = \prod_{i \sim j} \prod_{k,l} \left( P_{kl}^{w_{ij}^{(t)}} (1 - P_{kl})^{1 - w_{ij}^{(t)}} \right)^{z_{ik}^{(t)} z_{jl}^{(t)}}$

- Community transition probability

- $\Pr(Z^{(t)} | Z^{(t-1)}, A) = \prod_{i=1}^n \prod_{k,l} A_{kl}^{z_{ik}^{(t-1)} z_{il}^{(t)}}$

- Initial community assignment probability

- $\Pr(Z^{(1)} | \pi) = \prod_{i=1}^n \prod_k \pi_k^{z_{ik}^{(1)}}$

# Outline

---

- Motivation and background
- Dynamic stochastic block model
- Bayesian inference and algorithm
- Experimental results
- Conclusion

# Bayesian inference

- Conjugate prior for Bayesian inference

- Dirichlet:  $\Pr(\pi) = \frac{\Gamma(\sum_k \gamma_k)}{\prod_k \Gamma(\gamma_k)} \prod_k \pi_k^{\gamma_k - 1}$

- Beta:  $\Pr(P) = \prod_{k,l \geq k} \frac{\Gamma(\alpha_{kl} + \beta_{kl})}{\Gamma(\alpha_{kl})\Gamma(\beta_{kl})} P_{kl}^{\alpha_{kl} - 1} (1 - P_{kl})^{\beta_{kl} - 1}$

- Dirichlet:  $\Pr(A) = \prod_k \frac{\Gamma(\sum_l \mu_{kl})}{\prod_l \Gamma(\mu_{kl})} \prod_l A_{kl}^{\mu_{kl} - 1}$

- Joint probability

- $\Pr(\mathcal{W}_T, \mathcal{Z}_T) = \int \Pr(\mathcal{W}_T, \mathcal{Z}_T | \theta) \Pr(\theta) d\theta \propto$   
 $\prod_k \Gamma(n_k^{(1)} + \gamma_k) \prod_k \frac{\prod_l \Gamma(n_{k \rightarrow l}^{(1:T)} + \mu_{kl})}{\Gamma(n_{k \rightarrow \cdot}^{(1:T)} + \sum_l \mu_{kl})} \times$   
 $\prod_{k,l > k} B\left(\hat{n}_{kl}^{(1:T)} + \alpha_{kl}, n_{kl}^{(1:T)} - \hat{n}_{kl}^{(1:T)} + \beta_{kl}\right) \times$   
 $\prod_k B\left(\frac{\hat{n}_{kk}^{(1:T)}}{2} + \alpha_{kk}, \frac{n_{kk}^{(1:T)} - \hat{n}_{kk}^{(1:T)}}{2} + \beta_{kk}\right)$

# Bayesian inference — inference algorithm

---

- Gibbs sampling algorithm
- Two versions
  - offline: trying to fit the whole data
  - online: incrementally over time
- Taking advantage of data sparseness
- Linear (per iteration) in the size of the networks

# Outline

---

- Motivation and background
- Dynamic stochastic block model
- Bayesian inference and algorithm
- Experimental results
- Conclusion

# Experimental studies — data sets

---

- Synthetic
  - followed a procedure used in [NG03] and [LCZ+08]
  - 4 communities, 32 node each, with different noise levels
  - introduce random community transitions at each time stamp
  - ground truth on community membership is available
- Real-life blog data
  - 407 blogs, 148,681 links
  - 15 months of data partitioned into 9 time periods
  - two main types: blogs on politics and on technology
  - ground truth not available

# Experimental studies — metrics and baselines

## ● Metrics

- normalized mutual information (ground truth available)

$$MI(\mathcal{C}, \mathcal{C}') =$$

$$\sum_{C_i, C'_j} p(C_i, C'_j) \log \frac{p(C_i, C'_j)}{p(C_i)p(C'_j)} / \max(H(\mathcal{C}), H(\mathcal{C}'))$$

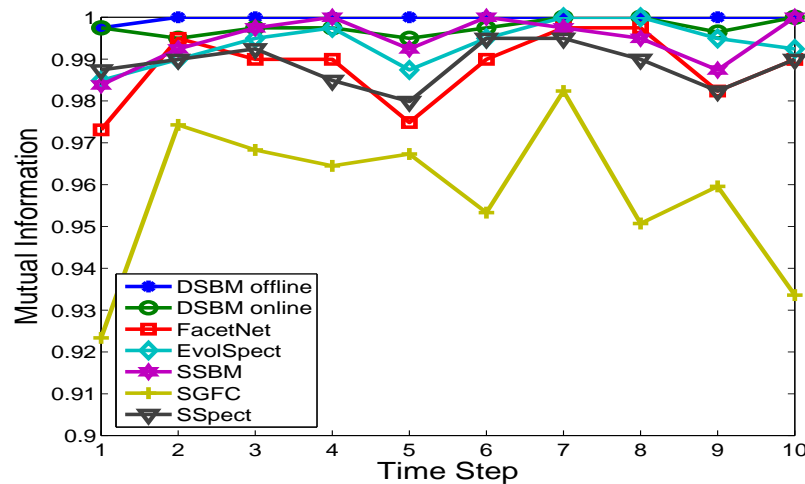
- modularity (ground truth unavailable)

$$Modu(\mathcal{C}) = \sum_k \left[ \frac{Cut(V_k, V_k)}{Cut(V, V)} - \left( \frac{Cut(V_k, V)}{Cut(V, V)} \right)^2 \right]$$

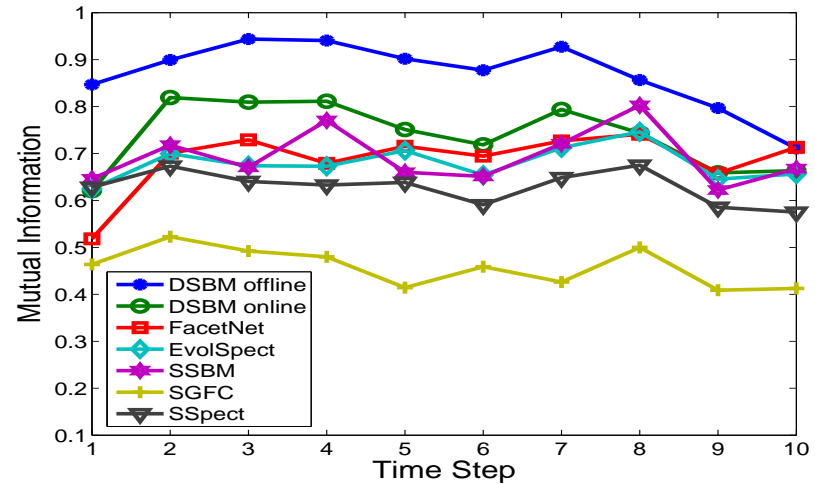
## ● Baselines

- static stochastic block model (SSBM) [YYT05]
- static spectral clustering (SSpect) [SM00]
- dynamic graph-factorization clustering (FacetNet) [LCZ+08]
- evolutionary spectral clustering (EvolSpect) [CSZ+07]

# Performance — synthetic data sets



(a) noise of level 1



(b) noise of level 3

Figure 2: The normalized mutual information with respect to the ground truth over the 10 time steps, of all the algorithms on the three data sets with different noise levels.

# Performance — convergence rate

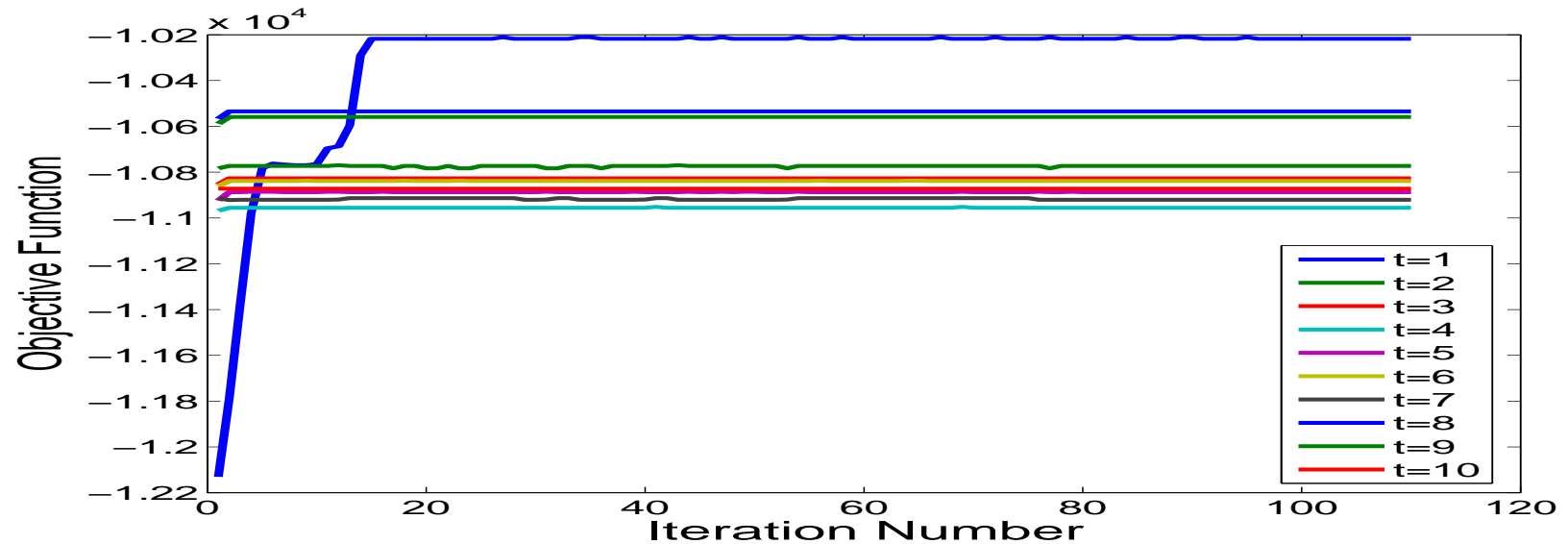
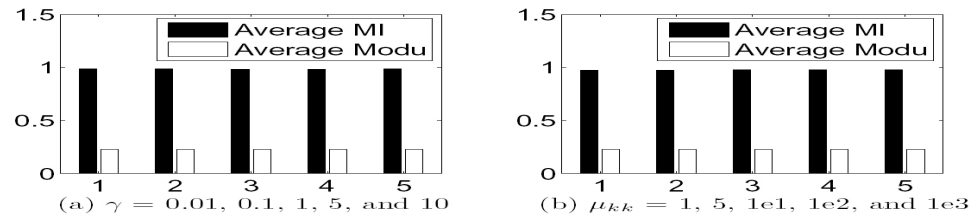
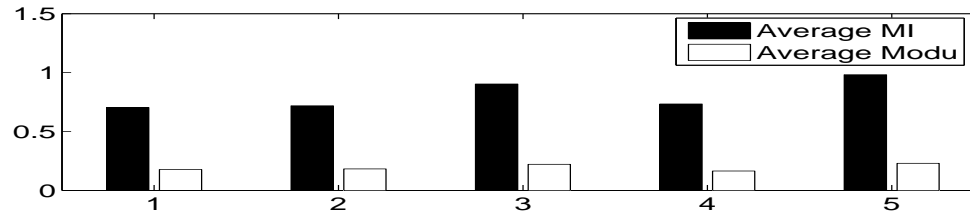


Figure 3: Convergence rate of Gibbs sampling procedure in the online learning.

# Performance — sensitivity to hyperparameters



(a)



(b)

Figure 4: The performance, in terms of the average normalized mutual information and the average modularity over all time steps, under different hyperparameters — (a)  $\gamma, \mu$  and (b)  $\alpha, \beta$ .

# Performance — blog data set

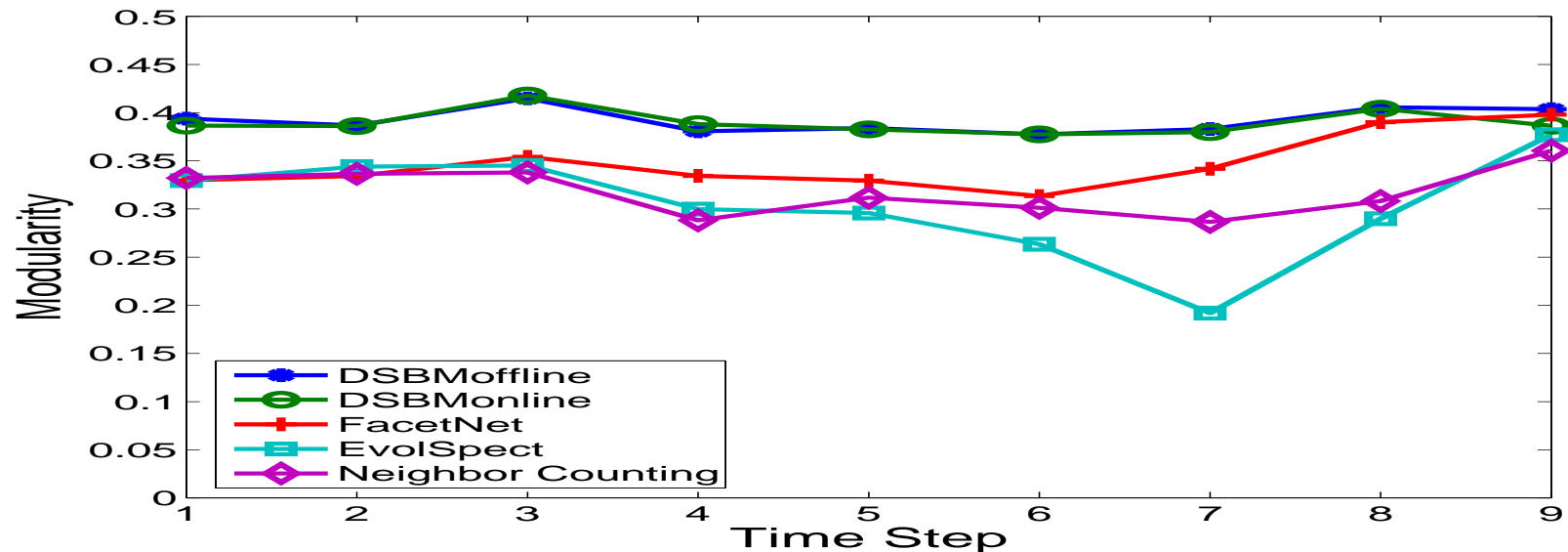


Figure 5: The performance, in terms of the modularity, of different algorithms (including the naive method using neighbor counting) on the NEC blog data sets.

# Performance — blog data set

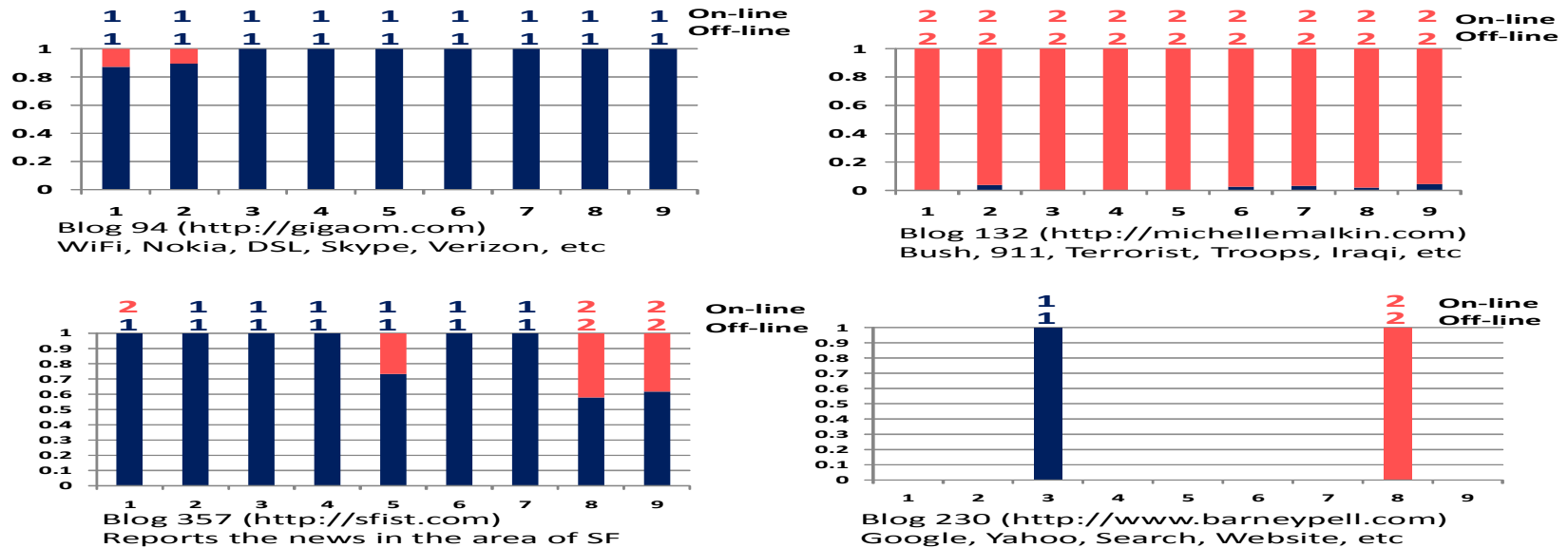


Figure 6: Neighbor distributions of the four representative blogs, the community results of the offline and online versions of DSBM, and some top keywords occurred in the blogs.

# Outline

---

- Motivation and background
- Dynamic stochastic block model
- Bayesian inference and algorithm
- Experimental results
- Conclusion

# Conclusion

---

- A probabilistic generative model that unifies the communities and their evolutions in an intuitive and rigorous way
- A Bayesian treatment that gives robust prediction of community memberships
- An efficient algorithm that makes framework practical in real applications
- Encouraging experimental results

# References

- [CSZ<sup>+</sup>07] Y. Chi, X. Song, D. Zhou, K. Hino, and B. L. Tseng. Evolutionary spectral clustering by incorporating temporal smoothness. In *Proc. of the 13th ACM SIGKDD Conference*, 2007.
- [HL76] P. Holland and S. Leinhardt. Local structure in social networks. *Sociological Methodology*, 1976.
- [LCZ<sup>+</sup>08] Y.-R. Lin, Y. Chi, S. Zhu, H. Sundaram, and B. L. Tseng. FacetNet: a framework for analyzing communities and their evolutions in dynamic networks. In *Proc. of the 17th WWW Conference*, 2008.
- [NG03] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phy. Rev. E*, 69(2), 2003.
- [SM00] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(8), 2000.
- [YYT05] K. Yu, S. Yu, and V. Tresp. Soft clustering on graphs. In *NIPS*, 2005.

# Performance — precision and recall

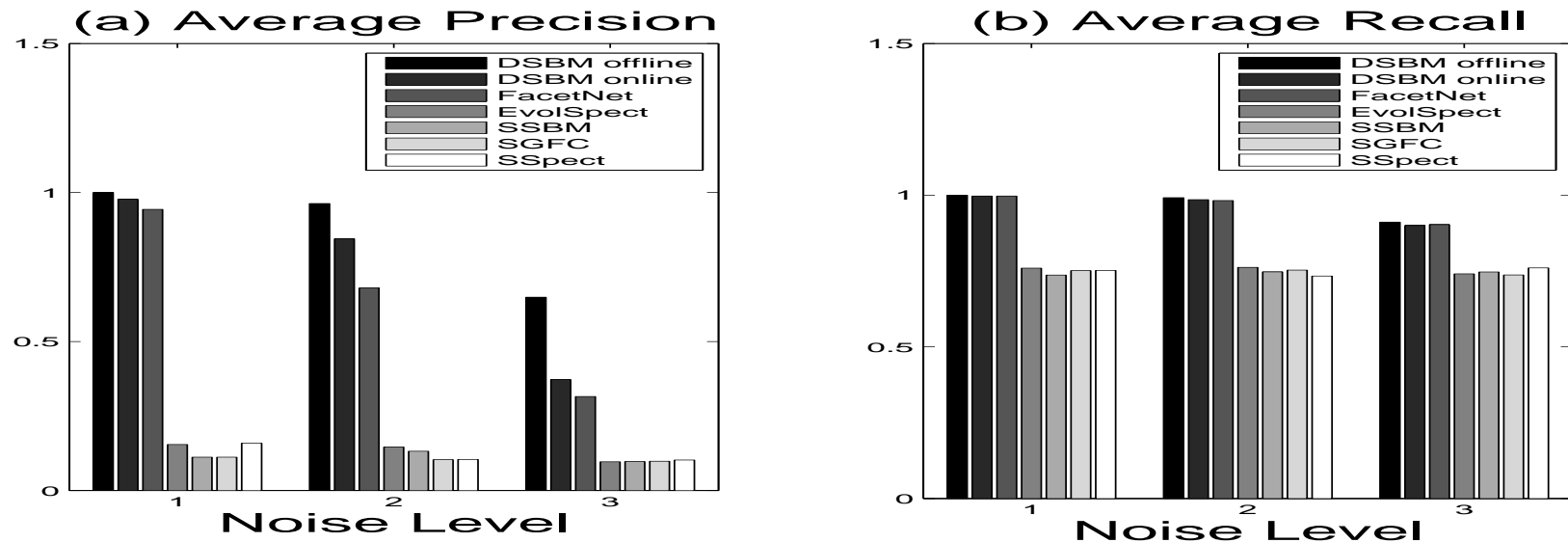


Figure 7: (a) The average precision and (b) the average recall over all the time steps for the three data sets with different noise levels.