TellTail: Fast Scoring and Detection of Dense Subgraphs

Bryan Hooi,[1] Kijung Shin,[2] Hemank Lamba,[3] Christos Faloutsos[3]
[1]National University of Singapore
[2]KAIST
[3]Carnegie Mellon University
bhooi@comp.nus.edu.sg, kijungs@kaist.ac.kr, hlamba@cs.cmu.edu, christos@cs.cmu.edu

Proof of Safe Pruning (Theorem 1)

Let $S^* = \arg\max_{S \subseteq V} \textsc{TellTail}(S)$ and $s = \lfloor n/2 \rfloor$. Define:

$$\Delta(S) = (s^\beta - (s-1)^\beta)\textsc{TellTail}(S) + \mu_0(s^\alpha - (s-1)^\alpha) \tag{1}$$

Theorem 1 (Safe Pruning)**.** *For any node i, if $i \in S^*$, we have:*

$$Dev_i \geq \Delta(S^*) \geq \Delta(S) \; \forall \; S \subseteq V \tag{2}$$

Thus, we can prune nodes with deviation less than $\Delta(S)$, for **any** *S.*

Proof. Define $k_* = |S^*|$, and $\gamma = k_*^\alpha - (k_* - 1)^\alpha$, and $\delta = k_*^\beta - (k_* - 1)^\beta$.

Since Dev_i is the sum of positive modularity terms along the ith row of B, Dev_i is an upper bound for how much adjusted mass the ith row can contribute to $\tilde{e}(S^*)$. Hence:

$$\textsc{TellTail}(S^*) \geq \textsc{TellTail}(S^* \setminus \{i\}) \tag{3}$$

$$\implies \frac{\tilde{e}(S^*) - \mu_0 \cdot k_*^\alpha}{k_*^\beta} \geq \frac{\tilde{e}(S^*) - \text{Dev}_i - \mu_0 \cdot (k_* - 1)^\alpha}{(k_* - 1)^\beta} \tag{4}$$

$$\implies \frac{\tilde{e}(S^*) - \mu_0 \cdot k_*^\alpha}{k_*^\beta} \geq \frac{\tilde{e}(S^*) - \text{Dev}_i - \mu_0 \cdot k_*^\alpha + \mu_0 \cdot \gamma}{k_*^\beta - \delta} \tag{5}$$

$$\implies -\delta \cdot \tilde{e}(S^*) + \mu_0 \cdot \delta \cdot k_*^\alpha \geq -k_*^\beta \cdot \text{Dev}_i + k_*^\beta \cdot \mu_0 \cdot \gamma \tag{6}$$

$$\implies \textsc{TellTail}(S^*) \leq \frac{\text{Dev}_i - \mu_0 \gamma}{\delta} \tag{7}$$

$$\implies \text{Dev}_i \geq \delta \cdot \textsc{TellTail}(S^*) + \mu_0 \cdot \gamma \tag{8}$$

$$\implies \text{Dev}_i \geq \Delta(S^*) \tag{9}$$

The second inequality $\Delta(S^*) \geq \Delta(S) \; \forall \; S \subseteq V$ follows from $\textsc{TellTail}(S^*) \geq \textsc{TellTail}(S)$, since this implies that $\Delta(S^*) \geq \Delta(S) \; \forall \; S \subseteq V$. □

Proof of Consistency (Theorem 2)

Theorem 2 (Consistency)**.** *Let G be a graph drawn from $\mathcal{G}$, and fix k. Define any fixed $z > 0$, and let $y_n = -\sigma(\hat{\mu}_n)(1 - z)/\xi$. Then as $n \to \infty$, there exists a sequence*[1] *of $N_n \to \infty, \epsilon_n \to 0$ such that:*

$$\frac{1 - \hat{F}(\hat{\mu}_n + y_n)}{1 - F(\hat{\mu}_n + y_n)} \xrightarrow{P} 1$$

where $\xrightarrow{P}$ denotes convergence in probability.

Fix N and let $n \to \infty$. Consider an n-node graph $G \sim \mathcal{G}$ and random subsets $S_1, \dots, S_N$ of size k. For any i, j, the probability that S_i and S_j are completely disjoint is $(\frac{n-k}{n})^k \to 1$ as $n \to \infty$ (recall that k is fixed). Extending this to all of the pairs of i, j by union bound, the probability that all of $S_1, \dots, S_N$ are disjoint goes to 1 as $n \to \infty$. This implies that with high probability, $S_1, \dots, S_N$ are disjoint and hence are i.i.d. samples of size k from $\mathcal{G}$, and their masses (denoted by $m_1, \dots, m_N$) are i.i.d. samples from F. Note that the event in which $S_1, \dots, S_N$ are non-disjoint has probability converging to zero, and since the statement we want to prove is about convergence in probability, this event can be ignored.

At this point, $m_1, \dots, m_N$ is an i.i.d. sample from a distribution F, and our algorithm $\textsc{Tail}$ estimates a GP distribution using maximum likelihood from this sample. (Smith 1987) shows that under these conditions, the maximum likelihood procedure produces consistent estimators of the tail probabilities of the distribution F. Formally:

$$\frac{1 - \hat{F}(\widehat{\mu_n} + y_n)}{1 - F(\widehat{\mu_n} + y_n)} \xrightarrow{P} 1$$

i.e. $\hat{F}$ converges to F, measured in terms of relative error with respect to the CCDF $1 - F$.

TellTail: Estimator for μ_0

Let A be the adjacency matrix, d be the column vector of node degrees, $B = A - d \cdot d^T/(2m)$, and $s = \lfloor n/2 \rfloor$. Then

[1] $N_n, \epsilon_n, \hat{\mu}_n$ denote the original variables $(N, \epsilon, \hat{\mu})$ indexed over runs corresponding to different values of n.

we will show that the following provides a reasonable approximator for μ_0, based on a Central Limit Theorem-based approximation:

$$\hat{\mu_0} = s^{-\alpha}(p_2 S_1 + z_{1-\epsilon}(p_2 S_2 + p_3(S_3 - 2S_2) + p_4(S_1^2 + S_2 - S_3) - (p_2 S_1)^2)^{1/2}) \tag{10}$$

where: $S_1 = \sum_{i<j} B_{ij}$, $S_2 = \sum_{i<j} B_{ij}^2$, $S_3 = \sum_{i=1}^n (\sum_{j=1}^n B_{ij})^2, p_r = \prod_{j=1}^r (s-j+1)/\prod_{j=1}^r (n-j+1)$, and $z_{1-\epsilon}$ is the $(1-\epsilon)$-quantile of a standard normal distribution.

To show this, we first give a lemma:

Lemma 1. *The mean and variance of $\tilde{e}(S)$ are:*

$$\begin{aligned} \mathbb{E}(\tilde{e}(S)) &= p_2 S_1 \\ \mathrm{Var}(\tilde{e}(S)) &= p_2 S_2 + p_3(S_3 - 2S_2) \\ &\quad + p_4(S_1^2 + S_2 - S_3) - (p_2 S_1)^2 \end{aligned} \tag{11}$$

Proof. Define random variables $Z_{ij} = B_{ij} 1\{i \in S, j \in S\}$. Note that $\tilde{e}(S) = \sum_{i<j} Z_{ij}$. Substituting this into $\mathbb{E}(\tilde{e}(S))$ and $\mathrm{Var}(\tilde{e}(S))$ and expanding with further computation gives the result. □

We now derive our estimator (10) for μ_0. Consider subgraphs of size $s = \lfloor n/2 \rfloor$. For subgraphs S of this large size, $\tilde{e}(S)$ can be expressed as a sum over individual edges: $\tilde{e}(S) = \sum_{i<j} Z_{ij}$, which recalling the Central Limit Theorem, suggests approximating $\tilde{e}(S)$ with a normal distribution. From our initial definition of the GP distribution and parameters, $\mu(s)$ is the minimum value in the GP distribution's support. Since we threshold the data at its $(1-\epsilon)$-quantile, $\mu(s)$ is the $(1-\epsilon)$-quantile of the subgraph mass distribution. For $s = \lfloor n/2 \rfloor$, this distribution is approximately Gaussian, so the corresponding $(1-\epsilon)$-quantile is:

$$\mu(n/2) \approx \mathbb{E}(\tilde{e}(S)) + z_{1-\epsilon}\sqrt{\mathrm{Var}(\tilde{e}(S))} \tag{12}$$

The power-law plot for μ passes through the point $(s, \mu(s))$, so substituting into the Dense Subgraph Power Law, its intercept is $\mu_0 = \mu(s)/(s)^\alpha$. Combining this with (11) gives the final result.

Computing S_1, S_2, S_3

Recall that $S_1 = \sum_{i<j} B_{ij}$, $S_2 = \sum_{i<j} B_{ij}^2$, $S_3 = \sum_{i=1}^n (\sum_{j=1}^n B_{ij})^2$.

Computing S_1 to S_3 naively is $O(n^2)$, by this can be sped up to linear time using matrix operations.

Lemma 2. *S_1 to S_3 can be computed in $O(m)$ time.*

Proof. However, $B = A - d \cdot d^T/(2m)$, a sum of a sparse and a low rank matrix. This substitution allows us to rewrite the expressions for S_1 to S_3:

$$S_1 = \frac{\sum_{i=1}^n d_i^2}{4m} \tag{13}$$

$$S_2 = \sum_{i<j} A_{ij}^2 - \frac{d^T A d}{2m} + \frac{(\sum_{i=1}^n d_i^2)^2 - \sum_{i=1}^n d_i^4}{8m} \tag{14}$$

$$S_3 = \frac{\sum_{i=1}^n d_i^4}{4m^2} \tag{15}$$

The only terms that require matrix operations are $\sum_{i<j} A_{ij}^2$ and $d^T A d$. Both can be computed in $O(m)$ time using standard sparse matrix operations. □

Subgraph Mass Distribution Plots in Real Data

Figure 1 plots the empirical distribution of subgraph masses for all 8 of our original datasets as plotted in Table III of the paper, and all 5 of the Twitter subset graphs.

Proof of NP Completeness

In this section, we show that maximizing TELLTAIL is NP-complete. Let $\tilde{e}_G(S)$ denote the adjusted mass of subset S with respect to the graph G, i.e. $\tilde{e}_G(S) = e(S) - d(S)^2/(4m)$, all with respect to G. Define the following two problems:

Problem 1 (Modularity). *Given a graph G, does there exist a subset S of its nodes that such $\tilde{e}_G(S) > 0$?*

The NP-hardness of modularity maximization was first shown by (Brandes et al. 2007), though this differs from the formulation here. The NP-hardness of this exact formulation was shown by (Dinh, Li, and Thai 2015). They do this by reducing the known NP-hard PARTITION problem, of partitioning n integers $x_1, \cdots, x_n$ into two subsets of equal sum, to the MODULARITY problem. To do this, they show that given $x_1, \cdots, x_n$, we can construct a graph such that a subset S of positive modularity exists iff there exists a partition of $x_1, \cdots, x_n$ into two halves of equal sum, which completes the reduction and establishes the NP-hardness of MODULARITY.

The problem we are interested in is:

Problem 2 (TELLTAILPROB). *Given a graph G', does there exist a subset S of its nodes such that* $\text{TELLTAIL}_{G'}(S) > 0$*?*

We now show our main NP-completeness result:

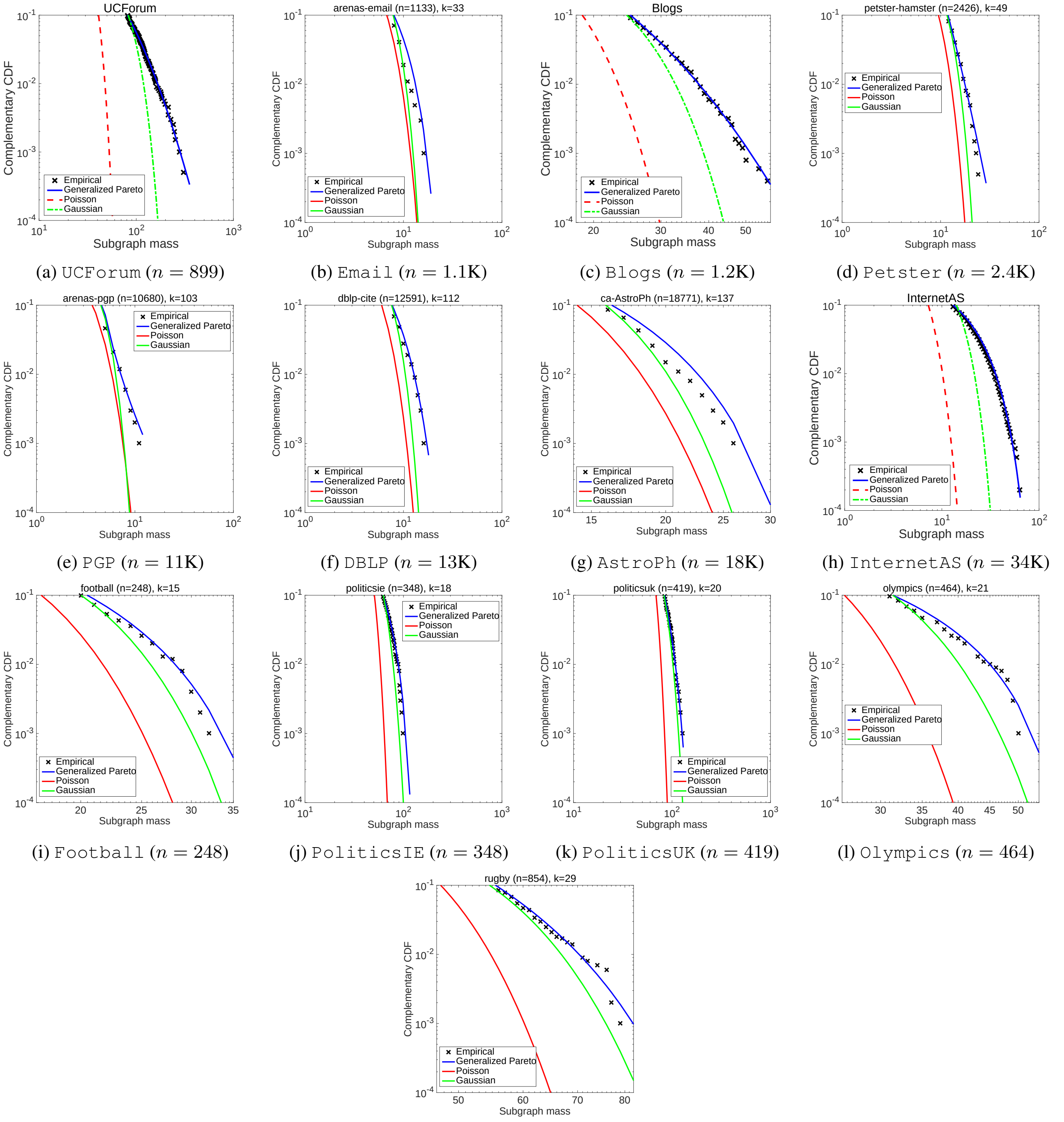
Theorem 3. TELLTAILPROB *is NP-complete.*

Proof. Given a subset S of the nodes of G, we can compute TELLTAIL(S) in polynomial time. Thus, TELLTAILPROB can be *verified* in polynomial time, and is therefore in NP. It remains to establish that TELLTAILPROB is NP-hard.

We do this by reducing MODULARITY to TELLTAILPROB: given an algorithm A' that solves TELLTAILPROB, we show that it can be used as a subroutine in a polynomial-time algorithm A to solve MODULARITY. Then, since we know that MODULARITY is NP-hard, this would imply that TELLTAILPROB is NP-hard as well.

Consider an instance of MODULARITY with graph G. Construct a new graph G' by adding r extra nodes to G, in which the extra nodes have no edges attached to them. Note then that for any subset S of the nodes of G, the adjusted mass of S is the same when computed with respect to either G or G': this is because in the formula $\tilde{e}(S) = e(S) - d(S)^2/(4m)$, none of the terms ($e(S)$, $d(S)$ or m) differ between G and G'.

Consider Eq. (10) for μ_0. Each of the p_i is at most 1, and the S_i are all constant as we increase r, which we can verify



(a) UCForum ($n = 899$) (b) Email ($n = 1.1$K) (c) Blogs ($n = 1.2$K) (d) Petster ($n = 2.4$K)

(e) PGP ($n = 11$K) (f) DBLP ($n = 13$K) (g) AstroPh ($n = 18$K) (h) InternetAS ($n = 34$K)

(i) Football ($n = 248$) (j) PoliticsIE ($n = 348$) (k) PoliticsUK ($n = 419$) (l) Olympics ($n = 464$)

(m) Rugby ($n = 854$)

Figure 1: *The GP distribution fits the empirical distribution of subgraph masses much more closely than other distributions.* Black crosses indicate the empirical distribution of subgraph masses for subgraphs of size $k = \lfloor\sqrt{n}\rfloor$, in the form of its complementary CDF. The colored curves are the best fit GP, Poisson, and Gaussian to the empirical distribution. The Poisson curve is far to the left of the empirical distributions because Poisson distributions greatly underestimate the number of dense subgraphs that we should observe.

from Eq. (6) to (8) of the original paper. Then, since G' has $n+r$ nodes, we have from Eq. (10) that $\mu_0 \leq (\frac{n+r}{2})^{-\alpha} B$, for some $B > 0$ that does not depend on r.

Set $r > 2(4mn^{\alpha}B)^{(1/\alpha)}$. Then in G', we have:

$$\begin{aligned}\mu_0 &\leq \left(\frac{n+r}{2}\right)^{-\alpha} B \\ &< \left(\frac{r}{2}\right)^{-\alpha} B \\ &= (4mn^{\alpha}B)^{(1/\alpha)\cdot(-\alpha)} B \\ &= \frac{1}{4mn^{\alpha}}\end{aligned}$$

Define the algorithm $A(G)$ for MODULARITY that given G, constructs G', runs our subroutine for solving TELLTAILPROB on G', and outputs $A'(G')$. We claim that $A(G)$ correctly solves MODULARITY. To show this, we consider two cases:

- **case 1:** there exists S such that $\tilde{e}_G(S) > 0$. Then since $\tilde{e}_G(S)$ is a fraction with an integer in the numerator and a denominator of $4m$, we have $\tilde{e}_G(S) \geq 1/(4m)$. Then, recalling that $\tilde{e}_G(S) = \tilde{e}_{G'}(S)$,

$$\begin{aligned}\text{TELLTAIL}_{G'}(S) &= \frac{\tilde{e}_G(S) - \mu_0 |S|^{\alpha}}{|S|^{\beta}} \\ &\geq \frac{\frac{1}{4m} - \mu_0 |S|^{\alpha}}{|S|^{\beta}} \\ &\geq \frac{\frac{1}{4m} - \frac{1}{4mn^{\alpha}} |S|^{\alpha}}{|S|^{\beta}} \\ &> 0\end{aligned}$$

  Thus $A(G)$ outputs the correct result in this case: $A'(G')$ will return 'true' since $\text{TELLTAIL}_{G'}(S) > 0$, so $A(G)$ will return 'true,' which is correct since $\tilde{e}_G(S) > 0$.

- **case 2:** there does not exist such an S; i.e. $\tilde{e}_G(S) \leq 0$ for all S. Then for all S,

$$\text{TELLTAIL}_{G'}(S) = \frac{\tilde{e}_G(S) - \mu_0 |S|^{\alpha}}{|S|^{\beta}} \leq \frac{\tilde{e}_G(S)}{|S|^{\beta}} \leq 0$$

  Thus $A(G)$ returns 'false', which is the correct result in this case as well.

In conclusion, $A(G)$ is a correct, polynomial time algorithm for MODULARITY, assuming we have a subroutine A' that solves TELLTAILPROB. Since MODULARITY is NP-hard by (Dinh, Li, and Thai 2015), this implies that TELLTAILPROB is NP-hard as well. Since we also showed that TELLTAILPROB is in NP, this implies that it is NP-complete. □

Accuracy Plots on Synthetic Data

For space reasons, accuracy results for identifying injected subgraphs on synthetic data was omitted from the main paper, and are included here in Figure 2.

Additional Results for Baseline Methods

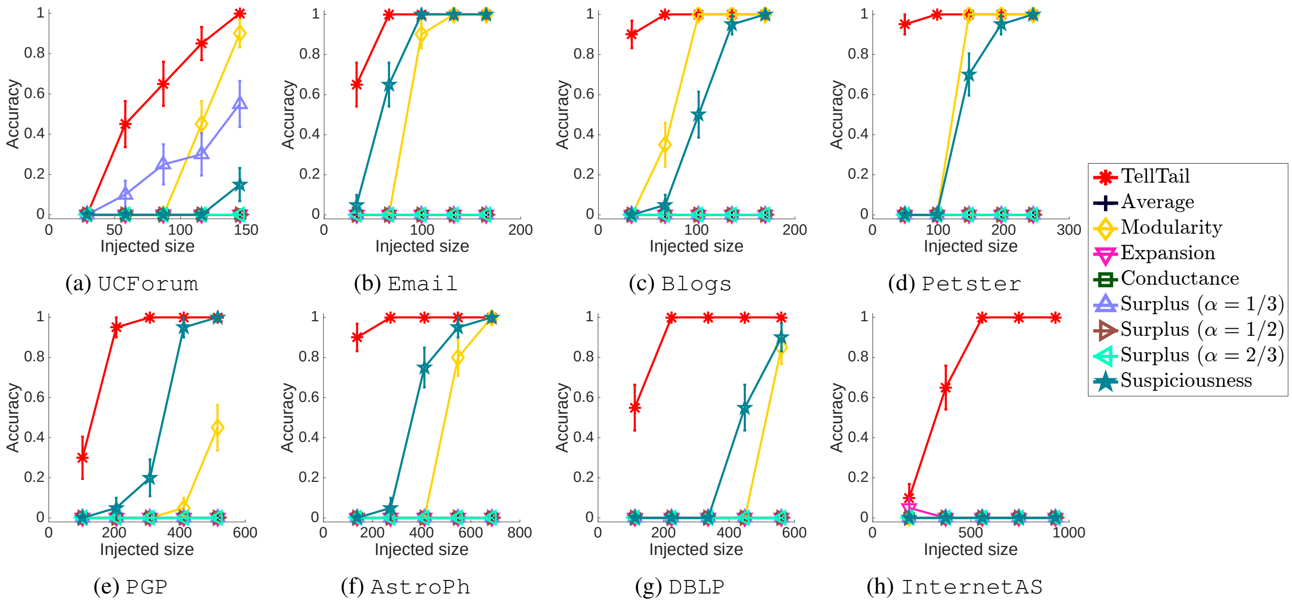
For space reasons, results varying α for the Edge Surplus method were omitted for the main paper, and are included here as Tables 1 and 2.

References

Brandes, U.; Delling, D.; Gaertler, M.; Görke, R.; Hoefer, M.; Nikoloski, Z.; and Wagner, D. 2007. On finding graph clusterings with maximum modularity. In *International Workshop on Graph-Theoretic Concepts in Computer Science*, 121–132. Springer.

Dinh, T. N.; Li, X.; and Thai, M. T. 2015. Network clustering via maximizing modularity: Approximation algorithms and theoretical limits. In *Data Mining (ICDM), 2015 IEEE International Conference on*, 101–110. IEEE.

Smith, R. L. 1987. Estimating tails of probability distributions. *The annals of Statistics* 1174–1207.



(a) UCForum (b) Email (c) Blogs (d) Petster

(e) PGP (f) AstroPh (g) DBLP (h) InternetAS

Figure 2: **TailF outperforms baselines in accuracy** for identifying injected subgraphs.

	TailF	Average	Modularity	Expansion	Conduct.	Surp.(1/3)	Surp.(1/2)	Surp.(2/3)	Susp.
Football	**1.00**	0.00	0.96	0.71	0.55	0.99	**1.00**	**1.00**	0.68
PolIE	**1.00**	0.00	0.94	0.87	0.86	0.16	0.75	0.86	0.38
PolUK	**0.96**	0.46	0.87	0.54	0.87	0.46	0.71	0.87	0.29
Olympics	**1.00**	0.00	0.85	0.74	0.67	0.99	**1.00**	**1.00**	0.63
Rugby	**0.97**	0.00	0.85	0.90	0.80	**0.97**	0.94	0.94	0.55

Table 1: **TailF outperforms baselines** in identifying ground truth communities.

	TellTail	TellTail+	Average	Modularity	Expansion	Conduct.	Surp.(1/3)	Surp.(1/2)	Surp.(2/3)	Susp.
Football	**0.78**	0.67	0.11	0.19	0.06	0.11	0.13	0.15	0.20	0.13
PoliticsIE	**0.90**	0.82	0.42	0.39	0.02	0.41	0.42	0.41	0.40	0.39
PoliticsUK	0.82	**0.84**	0.81	0.75	0.01	0.36	0.51	0.53	0.81	0.76
Olympics	0.68	**0.72**	0.20	0.21	0.11	0.15	0.22	0.29	0.44	0.22
Rugby	0.47	**0.49**	0.43	0.24	0.36	0.17	0.25	0.03	0.03	0.28

Table 2: **TellTail and TellTail+ outperform baselines** in identifying ground truth communities.