Dear Editor,

On behalf of all co-authors I enclose the revised manuscript entitled “Fast Analytical Methods for Finding Significant Labeled Graph Motifs”. We would like to thank the reviewers for their efforts and for their constructive, extensive, and insightful comments. We have restructured the paper according to their comments. We have highlighted the modifications in red in the revised manuscript. Below we give detailed point-to-point responses.

Best Regards,

Alfredo Pulvirenti

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Reviewer #1:**

**Reviewer question**

(1) The expression of using "colored graphs" is a bit confusing and definitely has nothing to deal with the well-known graph coloring problem. Also, because it is more common to refer to graphs with labels on the nodes/links as labeled/attributed graphs, the authors should clarify this in the paper.

**Authors Answer**

**We clarified the expression through the paper, we refer to “labeled” graphs instead of “colored” graphs. For this reason we also changed the title of the paper.**

**Reviewer question**

(2) Another confusing terminology is using EDD to refer to the expected degree random models, which is typically known as Chung-Lu model.

**Authors Answer**

**We completely reviewed the terminology used through the paper. We made it clearer that the Expected Degree distribution is the Chung-Lu model.**

**Reviewer question**

(3) The related work discussion is insufficient and does not cover the literature enough. For example, the literature of motif/graphlet counting has advanced so much to handle billion node/edge graphs extremely fast. None of these works is discussed in the paper (despite of how relevant these works to the paper). Please search "parallel graphlet counting". The proposed approach is currently taking 39 hours for counting 4-cliques in a network with ~54M edges. The runtime is too high (compared to the state of the art).

**Authors Answer**

**We thank the reviewer for this useful comment. We extended the related work accordingly, adding a discussion on state-of-the art algorithms for the subgraph enumeration. In particular, we mentioned that the approach using graphlet distribution is the fastest method for enumerating subgraphs of 3 and 4 nodes. The section has been extended with the following paragraph.**

***“The motif finding problem has attracted a lot of research concerning the design of efficient algorithms for the enumeration of subgraphs. Several methods have been proposed for the identification of induced and non-induced motifs [14, 18-20] of any size. Many of the tools are capable of dealing with motifs up to k= 9 nodes on medium size networks. When the size of the motif is small (k=3,4), the usage of graphlet decomposition techniques [21] has been proved to be the most efficient solution for unlabeled graphs even with large networks having billions of nodes and edges, because it lends itself to parallelism.”***

**Concerning the running time, according to our knowledge, the algorithm we used for counting the cliques, GLabTrie, is actually the most efficient solution for finding labeled motifs, while the approach using graphlet distribution is only for unlabeled motifs as we indicated in the added paragraph.**

**Reviewer question**

Also, the advantage of using analytical methods is to compute things extremely fast, why does the proposed approach take 39 hours if it is analytical? What is the complexity of this approach?

**Authors Answer**

**We clarified that 39 hours include both the time to count the 4-cliques and the time to compute the analytical p-values, even though, in practice, almost all the time is spent to count the cliques and the analytical p-value computation requires few seconds.**

**Reviewer question**

In addition, there is no mention to any of the existing papers on network hypothesis testing and attributed graph generation models.

**Authors Answer**

**We thank the reviewer for the suggestion. We added the following paragraph discussing attributed graph generation models:**

***“There has been some work to generate networks with real-world structural properties and correlated labels. In [22] the authors introduced the Attributed Graph Model (AGM), which exploits label correlations in connection to generative network models to jointly model network topology and node labels. In [23], the authors propose a generative model for labeled graphs called Multiplicative Attribute Graph (MAG) model. MAG generates the network taking into account the number of vertices, a set of prior probabilities for vertex label values and a set of affinity matrices specifying the probability of edges conditioned on the vertex labels. In [24], the authors describe AGWAN (Attribute Graphs: Weighted and Numeric), a generative model for random graphs with discrete labels and weighted edges.”***

**Reviewer question**

What's the impact of using Expected degree random model versus other models?

**Authors Answer**

**In the conclusion, we added a paragraph highlighting some advantages and limitations of the Expected Degree Distribution random model.**

**Reviewer question**

(4) The paper organization is extremely confusing. For example, many of the important materials (such as the baseline method, details of the analytical method, and results) are moved to the supplementary material. This makes the paper hard to read. The paper should at least have a brief description of the baseline method.

**Authors Answer**

**We deeply restructured the paper and the supplementary materials accordingly. In particular, we gave a more comprehensive discussion of the baseline method. Furthermore we added a deeper description of the EDD on unlabeled graphs.**

**Reviewer question**

(5) The notation needs careful revision. For example, using "Deg" to refer to the degree distribution is a bit confusing and makes the math hard to read. Why not use a symbol (e.g, $f\_d$)?

**Authors Answer**

**We modified the notation accordingly. We renamed the Random Variable Deg as $f\_D$.**

**Reviewer question**

(6) The network datasets used are extremely small. The authors should show their method on larger graphs.

**Authors Answer**

**We added a large citation network of 9 million nodes and 71 million edges to our dataset to better test the scalability of our framework. We show that the analytical model is very fast also in this case, requiring from few seconds to few hours to compute analytical p-values for all possible colored motifs with 3 and 4 nodes.**

**Reviewer question**

(7) The fonts in the Figures 5-6 are too small and unreadable.

**Authors Answer**

**Figures have been recreated using higher font sizes.**

**Reviewer #2:**

**Reviewer question**

Authors mostly focused on the bio-related venues, but motifs are also well studied in data mining/databases community. Induced motifs are also called as graphlets, and I recommend to include related papers on that topic. For example, "Graphlet decomposition: framework, algorithms, and applications, by Ahmed et al., KAIS journal (DOI: 10.1007/s10115-016-0965-5)" can be a starting point to explore literature in that direction. Also, most literature have focused on 3 and 4 vertex motifs, thus acknowledging the works on finding significant directed triangles or 4-vertex motifs would be great. Again, "Directed closure measures for networks with reciprocity, by Seshadhri et al., Journal of Complex Networks (DOI: 10.1093/comnet/cnv032)" is a good paper to start. Finally, Durak et al. (Degree Relations of Triangles in Real-world Networks and Models, CIKM'12) investigated the relation of vertex degrees in triangles and related to your EDD based method.

**Authors Answer**

**We thank the reviewer for these insightful suggestions and comments. We extended the Related Work section giving a discussion about algorithms for subgraph enumeration, including graphlet-based approaches. In the Conclusion section, we added a paragraph highlighting some advantages and limitations of the Expected Degree Distribution random model and investigated the relations between the model and the triangle counting distribution.**

**Reviewer question:**

Paper advances the previous state-of-the-art for colored motifs. However, EDD is not the best model for real-world networks, thus I'd love to see a discussion in the conclusion part about the thoughts and limitations for other more realistic graph models, like BTER (Community Structure and Scale-free Collections of Erdős-Rényi Graphs, Physical Review E 85(5):056109, May 2012, doi:10.1103/PhysRevE.85.056109).

**Authors Answer:**

**We thank the reviewer for these useful suggestions. In the Conclusion section, we added a paragraph highlighting advantages and limitation of the Expected Degree Distribution random model.**

**Reviewer question:**

Use case on DBLP is nice, but it'd be great to add an additional use case on PPI (or brain) network for which you can justify the results with ground-truth about certain protein interactions. I also wonder why do you need to generate 1000 random graphs for simulation. Could you do with less? Or what if you use more? There is a need for explanation there because the runtime comparison is directly related to that number.

**Authors Answer:**

**We added a yeast PPI network to our dataset with trusted and reliable interactions, which are the results of different validation processes as we explained in the Experimental section. We believe that this network could represent a valuable example of a network for which one can justify the results obtained, even though a real ground-truth about interactions is missing. Concerning the number of random graphs used for the simulation, we pointed out that most p-values are much lower or higher than the critical threshold (we chose 0.05), so they are little affected by the number of iterations. We also clarified that our choice of the number of random graph represents a good trade-off between the accuracy of the simulation-based p-value and the running time of the simulation-based approach.**

**Reviewer question:**

Paper is well written. There are a few typos I could find, better to proofread again:

- pg 3, ln 48: ED -> ER

- pg 4, ln 34: Title needs to be bold or as a subsection

- pg 5, ln 48: 'a injective' -> 'an injective'

- pg 8, ln 22: 'nodes degrees' -> 'node degrees'

- pg 9, ln 15: 'summing' -> 'multiplying'

- pg 22, ln 31: 'Familly' -> ‘Family'

**Authors Answer:**

**We fixed all the typos.**

**Reviewer question:**

Figures are too small, especially the text inside. Figs 1 and 2 need to be enlarged with readable text. Legend and axis titles/numbers are not readable in Figs 5 and 6.

**Authors Answer:**

**We recreated the figures with higher font sizes.**

**Reviewer #3:**

In this paper, the authors propose an efficient approach to detect motifs, which are defined as subgraphs that occur more frequently than expected, from colored graphs. The paper studies an important problem, provides a reasonable approach, conduct sound analysis, and show promising experimental results. The only concern is that the technical novelty with respect to existing work is not clear enough.

The topic of the paper belongs to "graph mining", and thus it is highly relevant to the disciplines of data mining and knowledge discovery.

Existing graph motif discovery work either detects multi-set colored motifs without considering the graph topology or detects uncolored subgraph motifs. The proposed method is the first one that can detect topological colored motifs.

The prior work is discussed. The current work differs from the authors' own previous work.

**Reviewer question:**

The proposed problem is different from the problems studied in previous work. In this sense, the previous research is presented clearly. However, the technical novelty with respect to existing work [14] is not clearly presented. It seems that the proposed model is an extension of [14] to the colored graph, and the analysis also largely follows the results in [14]. Therefore, a thorough comparison between the proposed work and [14] should be provided. What are the challenges when color is considered? What novel technical contributions this paper has made to deal with those challenges?

**Authors Answer:**

**The global setting of the analytical model follows the unlabeled case. However, for the labeled case we need to define new type of motifs and extend properly the Expected Degree distribution model. We have substantially restructured the paper to better highlight the aim of our approach together with the results we obtained.**

**Reviewer question:**

Figures and tables are necessary and sufficient, but one concern is that some curves and text in the figures are difficult to read. It would be better to make the font size larger and curve width larger.

**Authors Answer:**

**Figures have been recreated with higher font sizes and larger curve widths both in the main paper and in the Supplementary Material.**