ATLAS Note

Tagging Rate Function B-Tagging

A. Shibata
Department of Physics, Queen Mary
University of London
Mile End Road, London, UK

and

B. Clement
LPSC, Universite Joseph Fourier et INPG
CNRS/IN2P3, Grenoble, France

Abstract

W plus associated jet production forms a large background to numerous analysis to be studied at the LHC. Although recent developments in Monte Carlo generators enable generation of realistic events, event generation and detector simulation is limited by computing resources due to its large cross section. A method was developed to improve statistics from smaller samples of W plus jets when analysis requires the presence of b-tagged jets. In this paper, details of this technique and a thorough investigation of the results are presented within the context of top quark analyses.

1 Introduction

W plus associated jets is one of the major backgrounds to top physics processes. It is a challenging task to obtain the next-to-leading-order theoretical cross section for these processes due to the large amount of parton radiation though it has been accomplished up to W + 2 partons. Even though generation of these events using Monte Carlo generators can be done fairly efficiently, the amount of time taken to process samples with Geant4 full detector simulation is enormous. Currently it is only realistic to simulate about 1/100 of what is needed to match the integrated luminosity of the order 1 fb^{-1}. We have, however, developed a method to improve statistics from smaller amounts of data produced.

The reason this channel forms a large background is that its cross section is so large, that even b-tagging rejections of order one or two hundred will not be sufficient to filter these events out. Estimation of the W + light jets background therefore involves letting hundreds of events fail the b-tagging selection cuts to find out what fraction of the sample remains. One can improve on this by not throwing away the events without b-tagged jets, but by giving those events weights. This weight can be interpreted as the probability of the given event to contain mistagged jet(s). To achieve this, a parameterised tagging rate function, TRF (a function of η and p_T,) was used to calculate the event weight based on the kinematics of the jets found in each event.

As shown in detail in this paper, the histograms of kinematic distributions can be smoothed using this method. Smoothing of the distributions is particularly desirable when an analysis makes use of multivariate techniques. When distributions with large statistical fluctuations are used to train neural networks for example, they can easily become sensitive to these fluctuations rather than the real kinematic features. While smoothing can also be achieved by averaging techniques comparing the neighbouring bins in the same histogram, the TRF technique can be much more accurate since it extracts more information from the events which would otherwise be discarded.

This technique was also developed in the D0 experiment at the Tevatron and has successfully been used in the context of top physics analysis. This is the first attempt in the ATLAS experiment to incorporate this technique for b-tagging efficiency and its performance is investigated in detail in this paper.

2 Definition

A few terms used in the text need clear definition. These terms describe the performance of b-taggers.

- **Mistag Probability (also, Fake Rate)**: The probability that a given non-b jet is tagged. This varies as a function of jet p_T and η. It also depends on whether the jet contains c quark or tau lepton.
- **Rejection**: The inverse of mistag probability. A rejection of 200 means that one in 200 jets will be mistagged.
- **Efficiency**: The probability that a given heavy flavour jet is tagged.

Note that inverse of the mistag probability is the rejection on a per-jet basis. When considering more than one jet, the average mistag probability is not the average rejection.

The labelling of jets follows the definitions in [4]. A "Real b-jet" (or just "b-jet") is a jet with a true b quark with p_T >5 GeV in a cone of size $\Delta R = 0.3$ [1] around the jet direction. If a c quark or a tau lepton is found instead of a b quark, the jet is labelled as "c-jet" or "τ-jet" respectively. Other jets are labelled as "light jets" though the performance of b-tagging light jet rejection can be affected if there is a b/c/τ-jet close to the light jet. Therefore, light jets are labelled as "pure light jet" if there is no b, c quark or τ lepton within the cone of size $\Delta R = 0.8$.

3 Tagging Rate Function

Parameterisation of b-tagging as a function of p_T and $|\eta|$ was developed in the context of the fast detector simulation program, ATLFAST, as shown in [4] and we use the same parameterisation in this study. Such parameterisation was required since no track reconstruction is performed in fast simulation. Realistic estimation of rejection/mistag rate is crucial to the studies which depend on b-tagging. Two types of taggers were studied: IP2D (based on 2D impact parameter tagging, that is simple but with low rejection) and SV1+IP3D (based on secondary vertex and 3D impact parameter, with higher performance, also referred to as simply "SV1" in this

---

[1] $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$

text). Events from various physics processes were studied[4][2] using full detector simulation. The parameterisation was obtained by taking the average of the performance over different types of samples to cover the whole $\mathrm{p_T}$ -η phase space and to smear out effects caused by any particular generators. Figure 1 and 2 shows the rejection for the two taggers with the fixed b-tagging efficiency of 60%.

One can observe from these plots that IP2D has about an order of magnitude smaller rejection compared to IP3D+SV1. General features of the rejection as a function of p_T and $|\eta|$ are shared by the two: highest rejection is achieved in the central region ($|\eta| < 1.5$) for jets with p_T around 50GeV. High $|\eta|$, lower p_T and higher p_T regions all have lower rejection. The reason for this performance degradation is mainly due to track reconstruction inefficiencies. This was studied in [3] and is explained as follows:

- At large pseudo-rapidity, the particles cross more material and suffer more from multiple scattering leading to bad track reconstruction efficiency;
- Tagging of very high p_T jets is not very efficient since they contain many tracks in a small opening angle;
- Jets with low p_T contain low p_T tracks, more sensitive to multiple scattering.

4 Calculation of Jet and Event Weights

In the fast simulation of b-tagging, jets are tagged based on the fixed efficiency of 60%. This is the efficiency of the tagger used throughout the analysis. Real b-jets are tagged randomly at this rate. Other jets are classified into 4 types as defined in section 2 and a pseudo random number generator is used to decide whether a given jet is mistagged or not depending on the mistag rate for the given jet.

In the case of TRF weighting, event weights are calculated by taking the sum of the jet weights though this depends on the tag requirements: event probability for finding only one b-jet ("exclusive") in an event is different from event probability for finding one or more b-jets ("inclusive") in the event. One tag exclusive and inclusive probability can be calculated as:

---

[2] The parameterisation was produced with MC data simulated and reconstructed with ATHENA release 11. Due to the increase of inner detector material in detector geometry, performance degradation has been observed in release 12.

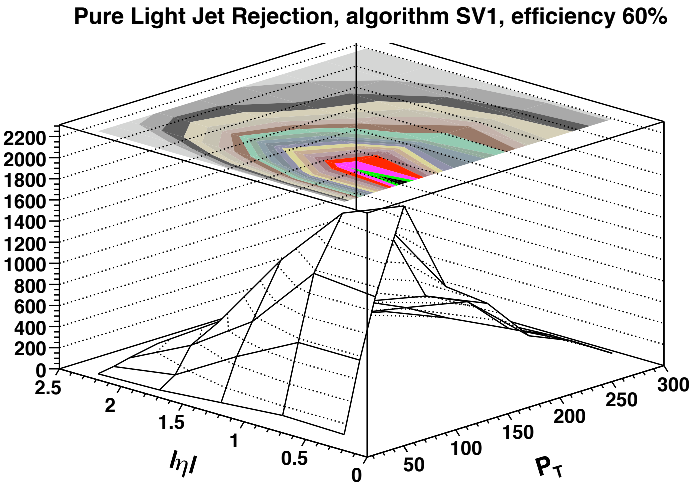

Figure 1: Rejection as a function of η and p_T for the algorithm combining SV1 and IP3D.

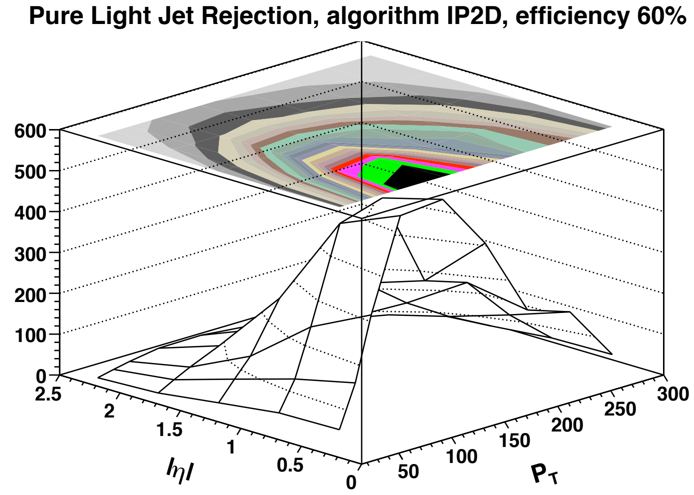

Figure 2: Rejection as a function of η and p_T for the algorithm using IP2D only.

$$
\begin{aligned}
P_{=1} &= \sum_{i=all\ jets} \{P(i) \times \prod_{j \neq i} (1 - P(j))\} && (1) \\
P_{\geq 1} &= 1 - \prod_{i=all\ jets} (1 - P(i)) && (2)
\end{aligned}
$$

where $P(i)$ is the probability of i-th jet (be it real b-jet or otherwise) to be tagged. With this, two tag inclusive probability can be calculated as:

$$P_{\geq 2} = P_{\geq 1} - P_{=1} \tag{3}$$

While event weights suffice for counting experiments, they do not specify which jet should be taken as b-tagged in case such selection is required for the purpose of the analysis (e.g. top reconstruction by adding W and b-tagged jet). Therefore, jets are tagged by the method similar to fast simulation using random numbers though this time under different requirements: exactly one jet will be tagged under the one tag exclusive requirement and two for the two tag requirement etc. For the two tag requirement, one considers all possible dijet pairs in the events and for each calculates the probability for both of these jets to be tagged. The probabilities will then be normalised so that the sum of the probabilities of all combinations is unity and a random number generator is used to pick one of the combinations according to their probabilities.

All the numbers calculated above are saved to the output file with the four vector information of jets and other objects and can be used when filling the histograms or when counting the number of events (by adding weights rather than adding the number of events).

5 Samples Used and Preselection

Table 1 shows the list of samples used in this analysis. The Alpgen [7] event generator was used to generate the events. The generator combines higher order matrix element calculations with soft parton shower radiation produced by an external generator, in this case Pythia [8] (MLM efficiency is calculated which should be used to scale the sample appropriately). This enables a much more realistic prediction of kinematic distributions compared to a pure parton-shower generator such as Pythia especially near the high-end tail region. The downside of Alpgen is that it can only calculate cross sections to leading-order since it does not consider any contributions from loop diagrams. To get more realistic estimates of cross sections, the next-to-leading order matrix element calculator, MCFM[2] was used to derive k-factors by taking the ratio of the next-to-leading order cross section and the leading-order cross section. The cross sections indicated in the table includes this factor when available (only up to W + 2 jets).

The last column in the table shows the corresponding integrated luminosity for the number of events generated for this study. The ATLAS detector aims to collect luminosity of the order $\sim 1\text{fb}^{-1}$ in the first year of operation and $\sim 10\text{fb}^{-1}$ in the second year. The computing resource required to generate such a large quantity of data with full detector simulation is a major challenge and the method developed in this study will clearly help the estimation of this background which has a very large cross sections compared to that for top events.

Process	Cross Sec. * BR (Alpgen × MLM Eff. × k-fac.)	Gen. Lumi.
W + 0 parton	$36833 \times 0.8351 \times 0.8 = 24607.39$ pb	13.86 pb^{-1}
W + 1 parton	$16621 \times 0.48 \times 0.861 = 6869.13$ pb	68.40 pb^{-1}
W + 2 parton	$8390 \times 0.3533.888 = 2632.20$ pb	85.39 pb^{-1}
W + 3 parton	$3787 \times 0.606 = 2294.92$ pb	24.06 pb^{-1}

Table 1: List of W+jets samples used for this study with their cross section and integrated luminosity corresponding to the data generated for this study.

Figure 3 shows the p_T and η of the jets of the above four samples. One can see that the jets are typically distributed in low p_T regions where rejection is lower. Samples with more parton radiation tend to have higher p_T jets since they are more likely to have hard radiation when more partons are radiated. Average jet rejection in these sample therefore ranges between ~ 100 (W + 0 parton) to ~ 400 (W + 3 parton).

In addition to the W+jets samples, $t\bar{t}$ and t-channel single top events were used to check the performance of the method in presence of true b quarks in the next section. MC@NLO [5] and AcerMC [6] generators were used for $t\bar{t}$ (not including the fully hadronic decay mode) and t-channel single top (not including the hadronic decay mode) events and their cross section and integrated luminosity for generated events are summarised in Table 2.

Generated events were passed through the ATLAS fast detector simulation and some basic selection cuts were applied including a 15 GeV cut on the p_T of the jets (reconstructed with the cone algorithm with radius 0.4 in ΔR) and 10 GeV on isolated electrons and muons.

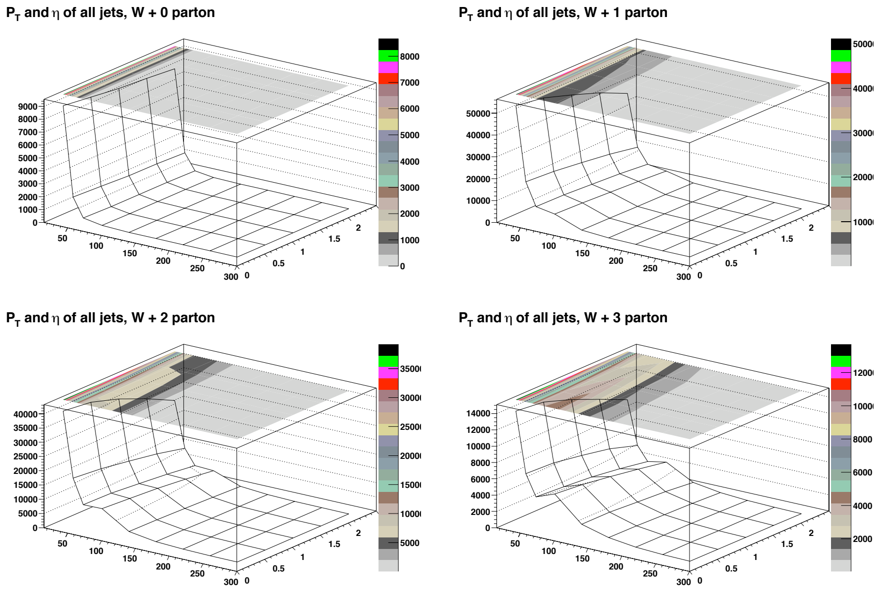

Figure 3: p_T and η distribution of all jets for each W+jets sample.

Process	Cross Section * BR	Generated Luminosity
$t\bar{t}$	461 pb (scaled to NLO+NLL [1])	1055.33 pb^{-1}
t-channel	81.3 pb	534.44 pb^{-1}

Table 2: List of top samples used in this study with their cross section and integrated luminosity corresponding to the data generated.

6 The Event Weight

Figure 4 and 5 show the distributions of the event weights calculated using TRF for different types of samples: $t\bar{t}$, t-channel single top and W plus jets. As the efficiency was set at the constant rate of 60%, the peaks at 60% indicate that there were jets with true b quarks. The weight distributions have some structure due to the combination of probabilities. We have attempted to explain some of the peaks. With $P(nbt|mb)$ standing for the probability of b-tagging n jets given m true b quarks, the following explains some of the outstanding structures of the plots:

1 excl:

- $P(1bt|3b) = 3 \times P(1bt|1b) \times P(0bt|1b) \times P(0bt|1b) = 28.8\%$

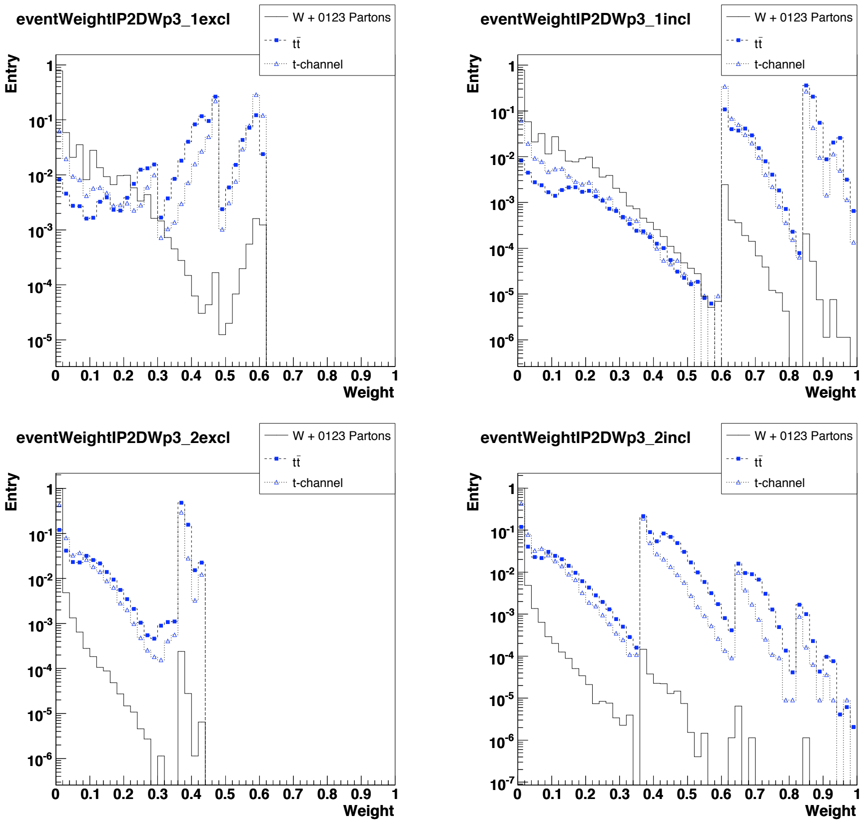

Figure 4: Event weights calculated using the TRF method with IP2D tagger (60% efficiency). One tag exclusive (left top), one tag inclusive (right top), two tag exclusive (left bottom) and two tag inclusive (right bottom). W+jets, t-channel single top and $t\bar{t}$ samples are compared.

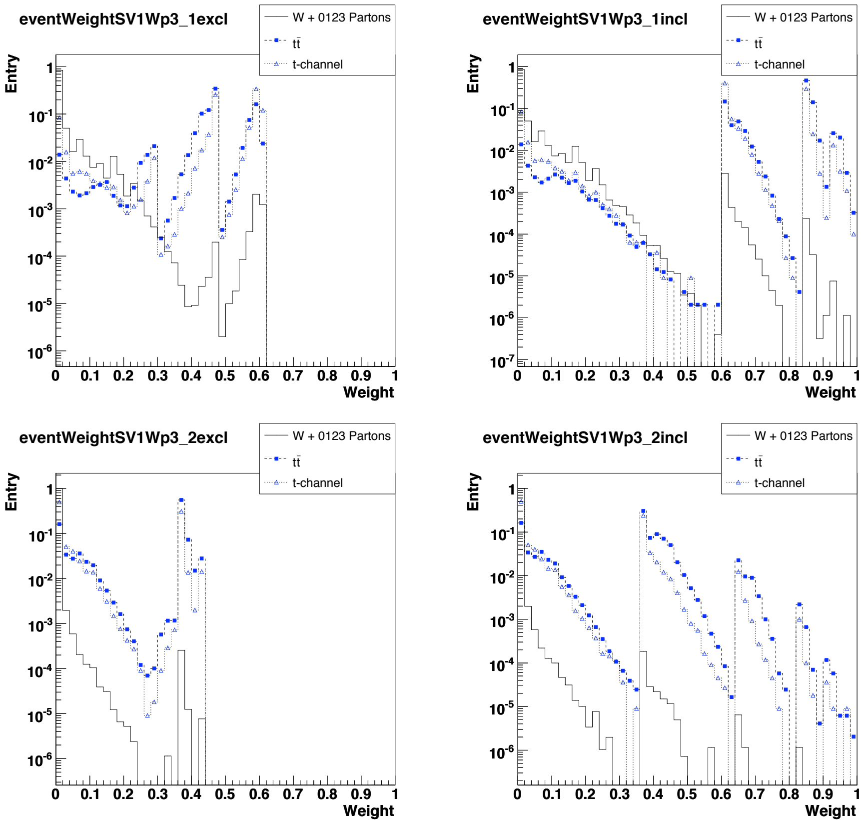

Figure 5: Event weights calculated using the TRF method with IP3D+SV1 tagger (60% efficiency). One tag exclusive (left top), one tag inclusive (right top), two tag exclusive (left bottom) and two tag inclusive (right bottom). W+jets, t-channel single top and $t\bar{t}$ samples are compared.

- $P(1bt|2b) = 2 \times P(1bt|1b) \times P(0bt|1b) = 48\%$
- $P(1bt|1b) = 60\%$

1 incl:

- $P(>= 1bt|1b) = 60\%$
- $P(>= 1bt|2b) = P(1bt|2b) + P(2bt|2b) = 84\%$
- $P(>= 1bt|3b) = P(1bt|3b) + P(2bt|3b) + P(3bt|3b) = 93.6\%$

2 excl:

- $P(2bt|3b) = 3 \times P(1bt|1b) \times P(1bt|1b) \times P(0bt|1b) = 43.2\%$
- $P(2bt|2b) = P(1bt|1b) \times P(1bt|1b) = 36\%$

Since $t\bar{t}$ and t-channel single top typically have one and two true b quarks respectively, the peaks are much more pronounced for these samples compared to W+jets where most of the jets are light jets. Tails on the left side (in exclusive plots) and the right side (in inclusive plots) of the peaks can be qualitatively understood as follows: The presence of extra jets tends to reduce the probability for exclusive requirements since all jets have tag probabilities between zero and one. With inclusive requirements, extra jets will always tend to increase the probability of events to have tagged jets.

7 Increase in Statistics - Case Study

To evaluate the increase in statistics, we studied simple but realistic event selection scenarios involving b-tagging requirements. The first analysis includes the following cuts:

(i) Missing transverse energy of the event to be greater than 20 GeV.

(ii) Number of leptons (electron or muon) to be equal to one.

(iii) Number of b-tagged jets to be more than or equal to one.

The efficiency of this selection and the number of events processed is shown in Table 3. W + 0, 1, 2, 3 parton samples are shown separately and the t-channel single top sample is also shown as a a reference (in t-channel single top there is one or two real b-jets). It shows that the sum of the weights of the events using TRF matches well with the number without TRF though there is a slight over estimation. The increase in the total number of events processed, "Gain", is of the order of 20 for all four samples.

Cut	W + 0p	W + 1p	W+2p	W + 3p	t-channel
Number of Evt	341075	469844	224764	55213	43450
Before TRF Weight					
$\not{E}_T$ Cut	63.84%	78.22%	81.06%	84.43%	90.74%
Lepton Num Cut	47.50%	48.88%	49.51%	49.94%	51.95%
Btag Num Cut	0.18%	1.24%	2.13%	3.43%	31.63%
Number After Cut	605	5831	4790	1893	13745
After TRF Weight					
Total Weight	2098.3	13240.5	10647.0	4024.51	26956.51
$\not{E}_T$ Cut	65.80%	78.21%	80.37%	84.14%	90.67%
Lepton Num Cut	30.22%	45.42%	47.18%	48.62%	52.10%
Btag Num Cut	30.22%	45.42%	47.18%	48.62%	52.10%
Weight After Cut	634.13	6012.93	5022.37	1956.80	14044.79
Number After Cut	11145	145032	95484	26450	21897
Gain	18.42	24.87	19.93	13.97	1.59

Table 3: The effect of the b-tag and and jet number requirements before and after using the event weight. There is no requirement on the total number of jets.

Changing the requirement to only one b-tag has little effect to the results here, of the order of 0.5%. For simplicity, and to understand the numbers better, consider the following analysis:

(i) Missing transverse energy of the event to be greater than 20 GeV.

(ii) Number of leptons (electron or muon) to be equal to one.

(iii) Number of b-tagged jets to be equal to one.

(iv) Total number of jets to be equal to one.

With the requirement of only one jet in the event, the event weight should be identical to the jet weight. The result is shown in Table 4. The gain changed non-uniformly over all samples up to about a factor of 2 but still of the same order as before.

As mentioned earlier, the event weight in this case is equal to the weight of the single jet found in the event. This was verified and for example in the W

Cut	W + 0p	W + 1p	W+2p	W + 3p	t-channel
Number of Evt	341075	469844	224764	55213	43450
Before TRF Weight					
$\not\!\!E_T$ Cut	63.84%	78.22%	81.06%	84.43%	90.74%
Lepton Num Cut	47.50%	48.88%	49.51%	49.94%	51.95%
Btag Num Cut	0.18%	1.23%	2.10%	3.30%	24.66%
Jet Num Cut	0.09%	0.94%	0.42%	0.10%	0.68%
Number After Cut	311	4411	953	54	295
After TRF Weight					
Total Weight	2098.3	13240.5	10647.0	4024.51	20522.14
$\not\!\!E_T$ Cut	65.86%	78.21%	80.36%	84.13%	90.65%
Lepton Num Cut	30.16%	45.53%	47.26%	48.67%	52.29%
Btag Num Cut	30.16%	45.53%	47.26%	48.67%	52.29%
Jet Num Cut	16.44%	34.46%	9.45%	1.44%	1.46%
Weight After Cut	340.04	4515.17	984.52	55.34	299.07
Number After Cut	7229	123847	29170	1645	745
Gain	23.24	28.08	30.61	30.46	2.53

Table 4: The effect of exactly one b-tag and and jet number requirement before and after using TRF event weight. This includes a requirement on the total number of jets.

+ 2 parton sample, the average weight of the jets in these events was 0.033. The inverse of this matches the gain in this sample. Notice that this number is not directly related to the average rejection. The average jet rejection in this sample for these event selection cuts was 339.

In the previous analysis, no additional cuts on jet p_T were performed though it is frequently the case that additional, harder, cuts are applied to objects for purification of the signal. Instead of the default 10 GeV and 15 GeV on leptons and jets respectively, we increased the cuts to 20 GeV and 30 GeV respectively. Note that the calculation of TRF was still done on the initial selection and additional cuts were applied afterwards. Table 5 summarises the result. As in the first analysis, the sum of the weights slightly overestimates the number of events without TRF though the difference is nearly within statistical uncertainties. Due to tighter cuts on jets and leptons, the number of events left after all cuts is significantly lower than before. The gain changed non-uniformly compared to the first analysis though it generally increased. This change in gain is caused by the loss of lower p_T jets. These jets typically have lower rejection.

Finally, we consider the two b-tag requirement. The other cuts are kept the same as before, though, this time event weight used is the 2 tag exclusive

Cut	W + 0p	W + 1p	W+2p	W + 3p	t-channel
Number of Evt	341075	469844	224764	55213	43450
Before TRF Weight					
$\not\!\!E_T$ Cut	63.84%	78.22%	81.06%	84.43%	90.74%
Lepton Num Cut	43.65%	41.41%	41.66%	41.78%	42.34%
Btag Num Cut	0.01%	0.25%	0.55%	1.12%	21.55%
Number After Cut	29	1160	1230	619	9365
After TRF Weight					
Total Weight	2098.3	13240.5	10647.0	4024.51	26956.51
$\not\!\!E_T$ Cut	65.80%	78.21%	80.37%	84.14%	90.67%
Lepton Num Cut	27.19%	39.14 %	40.50%	40.92%	42.77%
Btag Num Cut	1.48%	8.91%	12.31%	16.06%	35.74%
Weight After Cut	31.13	1179.15	1310.54	646.50	9633.92
Number After Cut	536	44774	32586	9546	14497
Gain	18.48	38.60	26.49	15.42	1.55

Table 5: The effect of the one or more b-tag requirement before and after using the TRF event weight. Additional p_T cut on jets and leptons are applied as described in the text.

weight and therefore the result with the same cuts can differ. The following is the list of cuts for this analysis:

(i) Missing transverse energy of the event to be greater than 20 GeV.

(ii) Number of lepton (electron or muon) to be equal to one.

(iii") Number of b-tagged jets to be equal to two.

No additional p_T cuts were applied unlike in the previous analysis.

As shown in Table 6 In this case, almost no events are left after the selection cuts and a gain of the order of $\sim$ 400 was obtained. The sum of the TRF weights after all the cuts is within the statistical uncertainty of the number of events left before applying weights showing consistency of the results. The gain here is of the order of magnitude of the square of the gain with a single tag requirement ($\sim$ 20) as expected.

8 Improvements in Distributions

In the previous section, it was shown that TRF tagging increases the number of events used by a factor of $\sim$ 20 for in the case of the one tag requirement

Cut	W + 0p	W + 1p	W+2p	W + 3p	t-channel
Number of Evt	341075	469844	224764	55213	43450
Before TRF Weight					
$\not\!\!E_T$ Cut	63.84%	78.22%	81.06%	84.43%	90.74%
Lepton Num Cut	47.50%	48.88%	49.51%	49.94%	51.95%
Btag Num Cut	0.00%	0.01%	0.03%	0.12%	6.66%
Number After Cut	7	40	70	69	2892
After TRF Weight					
Total Weight	28.96	134.90	222.38	4024.51	10598.19
$\not\!\!E_T$ Cut	61.45%	78.17%	81.26%	84.44%	90.76%
Lepton Num Cut	34.15%	34.81 %	42.98%	47.50%	51.67%
Btag Num Cut	34.15%	34.81%	42.98%	47.50%	51.67%
Weight After Cut	9.89	46.96	95.58	82.42	3096.14
Number After Cut	2284	13876	48318	21094	17887
Gain	326	346	690	306	6.18

Table 6: The effect of double b-tag requirement before and after using the event weight. No additional cuts on lepton and jet pt are applied.

and $\sim$ 400 in the case of the two tag requirement. The result of this is that the errors on the histograms from these samples become smaller and the final kinematic distributions are predicted more accurately. Otherwise spiky distributions are smoothed. These can now be used more reliably in the analysis using multivariate techniques since that would otherwise become more sensitive to statistical fluctuations than the general kinematic features.

Figure 6 compares the distributions without TRF tagging and with TRF tagging with the one tag requirement. The smoothing effect is clearly visible here in all variables. Figure 7 shows the same variables with the two tag requirement. The improvement is much larger since the statistical gain is an order of magnitude larger. These plots combine the four W+jets samples according to their luminosity and they are scaled to the overall integrated luminosity of 1 fb^{-1}.

The relative importance of this method in the context of t-channel analysis can be seen in figure 8. W + jets background is one of the largest backgrounds to this analysis along with $t\bar{t}$ channel and the effect of smoothing is very beneficial.

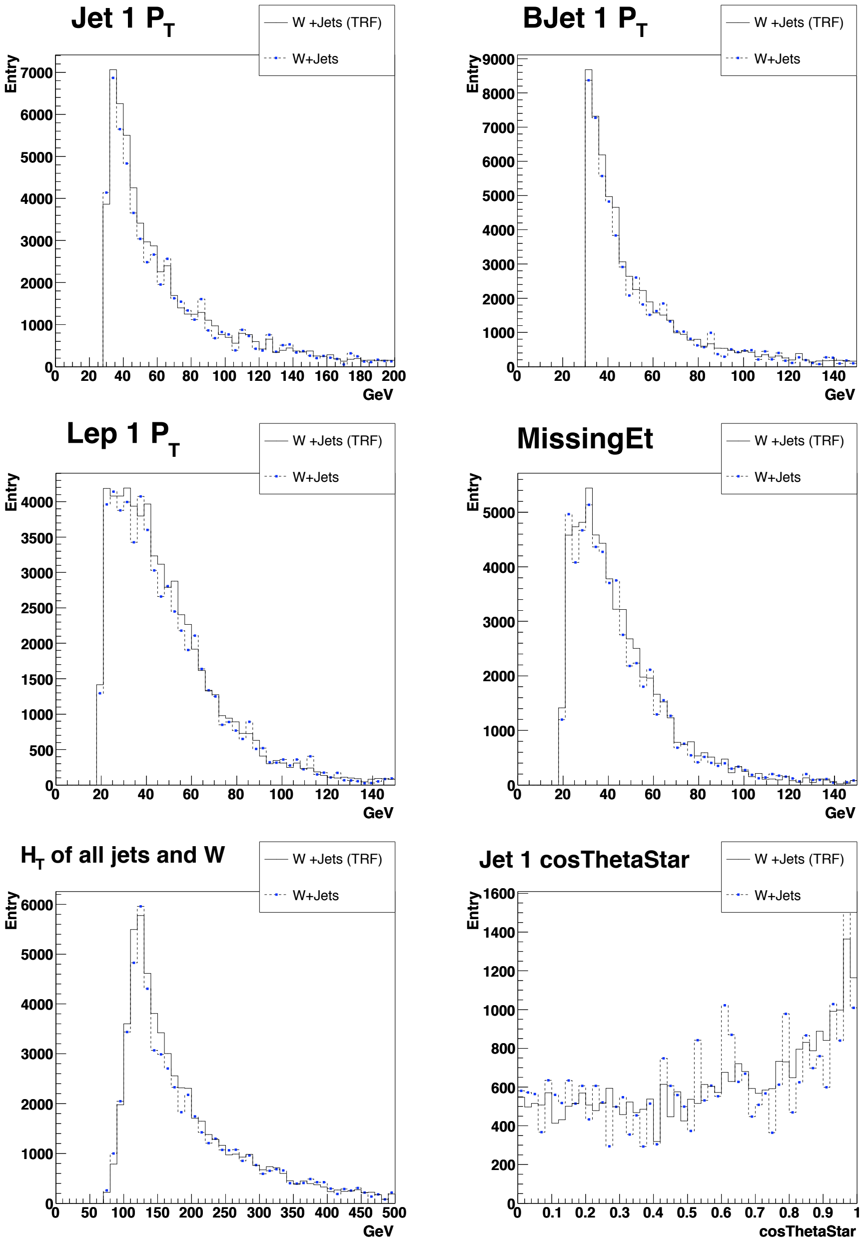

Figure 6: Comparison of the histograms of the event kinematics produced before (dotted line) and after (solid line) using TRF weight with the one b-tag requirement. Lepton and $\not{E}_T$ with additional p_T cuts were applied (see text). All four W+jets samples are added together with weights according to their cross section. From top left to bottom right; p_T of the leading jet, p_T of the b-tagged jet, p_T of the lepton, $\not{E}_T$, HT of all jets, the lepton and $\not{E}_T$, $cos(\theta)$ of leading jet in the event rest frame.

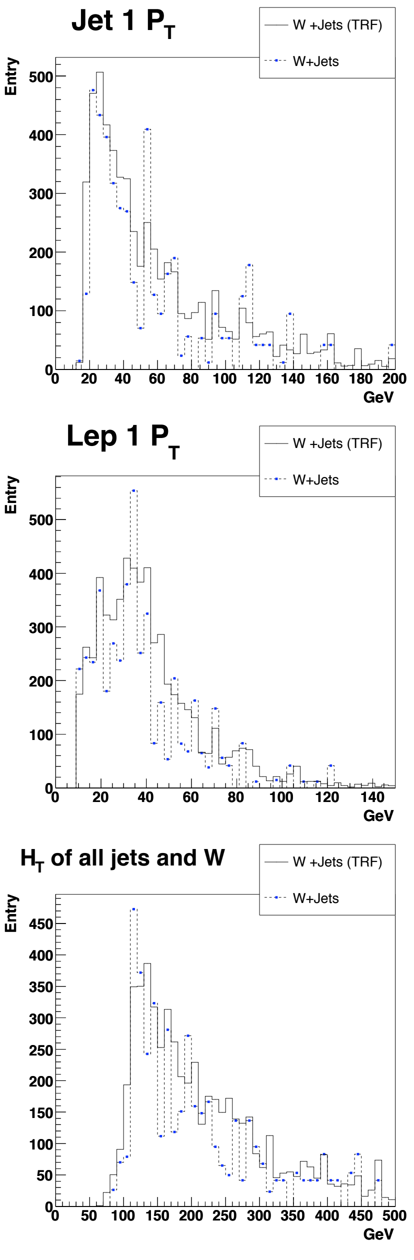

Figure 7: Same plots as Figure 6 with the two tag requirements. No additional p_T cuts were applied.

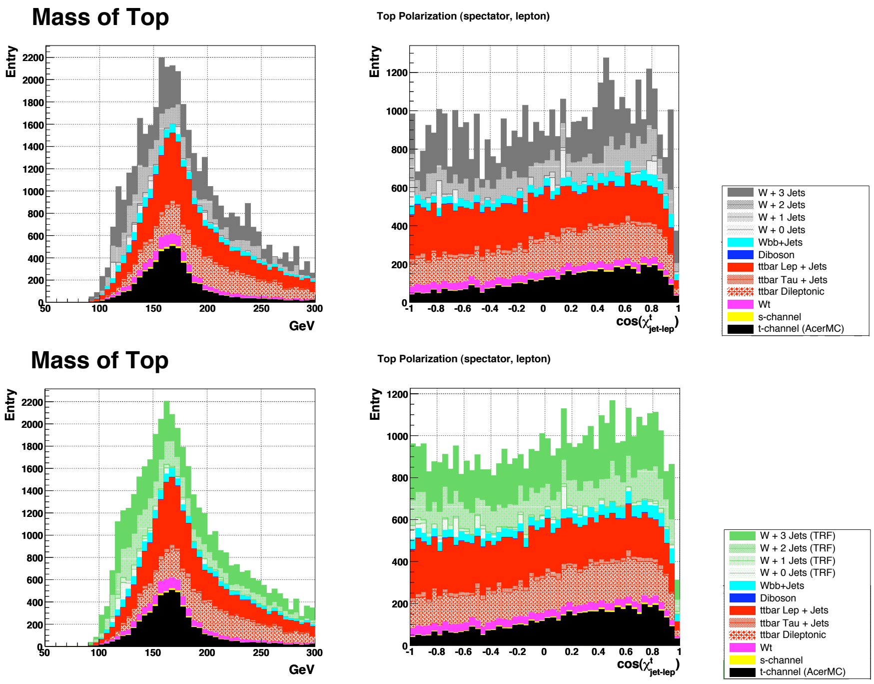

Figure 8: Comparison of the distributions of top mass (left) and top polarisation (right) before (top) and after (bottom) using TRF weight.

9 Acknowledgements

The authors would like to thank Arnaud Lucotte for encouraging support for this project. We would also like to thank Reinhard Schwienhorst for useful discussions. This work is largely based on the work by J. B. de Vivie to parameterise the b-tagging rejection. We are most grateful for his efforts and allowing us to use the parameterisation shown in figure 1 and 2.

References

[1] R. Bonciani et al. NLL resummation of the heavy-quark hadroproduction cross-section. *Nucl. Phys.*, page 424, 1998.

[2] J. Campbell and R. K. Ellis. *MCFM v5.0*, 2006.

[3] S. Correard et al. b-tagging with DC1 data. *ATLAS Notes*, ATL-PHYS(2004-006), 2003.

[4] Cavalli D. et al. Performance of the atlas fast simulation atlfast. *ATLAS Notes*, com-phys(012), 2007.

[5] F. Frixione and B. Webber. *The MC@NLO 3.2 Event Generator.*

[6] B. P. Kersevan and R. W. Elzbieta. The Monte Carlo Event Generator AcerMC version 3.4 with interfaces to PYTHIA 6.3, HERWIG 6.5 and ARIADNE 4.1. *arXiv*, hep-ph(0405247), 2004.

[7] L.M. Mangano. Alpgen, a generator for hard multiparton process in hadronic collisions. *arXiv*, hep-ph(0206293), 2002.

[8] T. Sjostrand, S. Mrenna, and P. Skands. Pythia 6.4: Physics and manual. *arXiv*, hep-ph(0603175), 2006.